

# Dynamic Rate and Channel Selection in Cognitive Radio Systems

R. Combes and A. Proutiere  
KTH, The Royal Institute of Technology

**Abstract**—In this paper, we investigate dynamic channel and rate selection in cognitive radio systems which exploit a large number of channels free from primary users. In such systems, transmitters may rapidly change the selected (channel, rate) pair to opportunistically learn and track the pair offering the highest throughput. We formulate the problem of sequential channel and rate selection as an online optimization problem, and show its equivalence to a *structured Multi-Armed Bandit* problem. The structure stems from inherent properties of the achieved throughput as a function of the selected channel and rate. We derive fundamental performance limits satisfied by *any* channel and rate adaptation algorithm, and propose algorithms that achieve (or approach) these limits. In turn, the proposed algorithms optimally exploit the inherent structure of the throughput. We illustrate the efficiency of our algorithms using both test-bed and simulation experiments, in both stationary and non-stationary radio environments. In stationary environments, the packet successful transmission probabilities at the various channel and rate pairs do not evolve over time, whereas in non-stationary environments, they may evolve. In practical scenarios, the proposed algorithms are able to track the best channel and rate quite accurately without the need of any explicit measurement and feedback of the quality of the various channels.

## I. INTRODUCTION

In cognitive radio systems, radio devices may access a potentially large number of frequency bands or channels. An example of such systems are those exploiting “white space” spectrum, the unused part of the TV/UHF spectrum (unallocated or not used locally). The FCC 2008 ruling allowed unlicensed devices to use parts of this spectrum, provided that devices can detect primary users (TV transmitters and wireless microphones). As a part of the 2010 ruling [1], FCC mandates the use of a geolocation database to identify which frequencies are free from primary users. By querying the geolocation database, we are guaranteed to obtain a set of channels free from primary transmitters and we avoid the difficult problem of sensing primary users.

We consider systems exploiting channels known to be free from primary users. For the transmission of each

packet, transmitters can select a coding rate from a finite predefined set (as in 802.11 systems for example) and a channel from the set of available channels. The outcome of a packet transmission is random, and the probabilities of successfully transmitting a packet using the various (channel, rate) pairs are a priori unknown at the transmitter; they need to be learnt based on trial and error. These probabilities can vary significantly and randomly over time and across channels; they also strongly depend on the chosen coding rate. As a consequence, tracking the best (channel, coding rate) pair for transmission may greatly improve the system performance. In this paper, we aim at designing sequential channel and coding rate selection schemes that efficiently track the best available channel and the corresponding coding rate.

As shown in previous works, see e.g. [2], [3], RSSI (Receive Signal Strength Indicator) is a poor predictor of channel quality, and hence of the packet successful transmission probabilities. In OFDM systems for example, this stems from the fact that RSSI does not report the individual signal strength experienced on the various sub-carriers. In order to accurately estimate the quality of a wide-band channel, more sophisticated techniques with specific hardware are needed [4], [5]. But these techniques are not typically supported in current commercial radio hardware. Instead, we need to infer the quality of each channel at each transmission rate through probing. Several packets have to be sent on each channel and at each rate to construct a reliable estimate of the channel quality. In the design of channel and rate selection schemes, we then face a classical exploration vs. exploitation trade-off problem. We need to exploit the (channel, rate) pair that has offered the best throughput so far, whilst constantly exploring other pairs in case one of them is actually optimal.

We rigorously formulate the design of the optimal sequential (channel, rate) selection algorithms as an online stochastic optimization problem. In this problem, the objective is to maximize the number of packets successfully sent over a finite time horizon. We show that this problem reduces to a Multi-Armed Bandit (MAB) problem [6]. In MAB problems, a decision maker

sequentially selects an action (also called an “arm”), and observes the corresponding reward. Rewards of a given arm are random variables with unknown distribution. The objective is to design sequential action selection strategies that maximize the expected reward over a given time horizon. These strategies have to achieve an optimal trade-off between exploitation (actions that have provided high rewards so far have to be selected) and exploration (sub-optimal actions have to be chosen so as to learn their average rewards). For our (channel, rate) selection problem, the various arms correspond to the decisions available at the transmitter to send packets, i.e., an arm corresponds to a channel and a coding rate. When a (channel, rate) pair is selected for a packet transmission, the reward is equal to 1 if the transmission is successful, and equal to 0 otherwise. The average successful packet transmission probabilities at the various (channel, rate) pairs are unknown, and have to be learnt.

The MAB problem corresponding to the design of channel and rate selection mechanisms is referred to as a *structured* MAB problem. It differs from classical MAB problems. (i) First, the rewards associated with the various rates on a given channel are stochastically correlated, i.e., the outcomes of transmissions at different rates are not independent: for example, if a transmission at a high rate is successful, it would be also successful at lower rates. (ii) Then, the average throughputs achieved at various rates exhibit natural structural properties. For a given channel, the throughput is a unimodal function of the selected rate. (iii) In addition, most often, on all channels, the packet successful transmission probabilities are close to 1 at low rates, and abruptly decrease to 0 as the rate increases. This additional structure, referred to as graphical unimodality, allows us to predict the outcomes of transmissions on various channels. As we demonstrate, correlations and (graphical) unimodality are instrumental in the design of channel and rate selection mechanisms, and can be exploited to learn and track the best (channel, rate) pair quickly and efficiently. Finally, note that most MAB problems consider stationary environments, which, for our problem, means that the successful packet transmission probabilities for the different (channel, rate) pairs do not vary over time. In practice, the transmitter faces a non-stationary environment as these probabilities could evolve over time. We consider both stationary and non-stationary radio environments.

In the case of stationary environments, we derive an upper bound of the expected reward that can be achieved in structured MAB problems. This provides a fundamental performance limit that *any* (channel, rate) selection algorithm cannot exceed. This limit quantifies

the inevitable performance loss due to the need to explore sub-optimal (channel, rate) pairs. It also indicates the performance gains that can be achieved by devising schemes that optimally exploit the correlations and the structural properties of the MAB problem. We present sequential (channel, rate) selection algorithms that optimally exploit the structural properties of the problem: for our algorithms, we prove that the performance loss due to the need to explore sub-optimal (channel, rate) pairs does not depend on the number of available rates. We also extend our algorithms to non-stationary radio environments. Finally, we evaluate the performance of the proposed algorithms using an office white-space testbed operating in the 500MHz-600MHz band, and simulation experiments.

#### *Contributions and paper organization.*

- The next section is devoted to the related work.
- In Sections III and IV, we present the models and objectives. We formulate the design of (channel, rate) selection algorithms as an online optimization problem, and establish its equivalence to a structured MAB problem.
- We derive, in Section V, a performance upper bound satisfied by any (channel, rate) selection algorithm, depending on the assumptions made on the structure of the problem – three scenarios with increasing structure are considered: 1. no structural assumption is made; 2. the throughput on each channel is a unimodal function of the rate; 3. the throughput is a graphically unimodal function of the channel and rate. We also quantify the performance gains that one may achieve by exploiting the structural properties of the problem.
- In Section VI, we propose three (channel, rate) selection algorithms, one for each of the above scenarios, and analyze their performance in stationary radio environments. We prove that our algorithms optimally exploit the structural properties of the throughputs.
- The extensions of our algorithms to non-stationary radio environments are briefly presented in Section VII.
- Finally, Section VIII is devoted to test-bed and simulation experiments.

## II. RELATED WORK

First observe that the joint channel and rate selection problem is considerably more difficult than detecting channels with no primary users as considered in a lot of recent works, see e.g. [7], [8], [9], [10], [11], [12], [13]. In some of these papers, a MAB framework has

been used to design primary user detection algorithms. The presence or the absence of primary users just means that a channel is either good or bad. When selecting both channel and rate, the dimension of the problem becomes larger, and there are multiple and numerous possible channel states. Primary users are not considered in our work, as we assume that transmitters can use a geolocation database to get a list of channels free from primary users [1].

It should also be observed that most of the work on dynamic spectrum access considers stationary radio environments. In [9], [10] for example, the authors use classical stochastic control techniques (Markov Decision Processes) to sequentially select a channel for transmission. The underlying assumption is that the environment is stationary, i.e., the packet successful transmission probabilities do not evolve over time. In this paper, both stationary and non-stationary radio environments are explored. Test-bed experiments actually suggest that the environment is non-stationary in practice, even in networks where nodes do not move such as indoor offices, see e.g. [3].

Our problem resembles the rate adaptation problem in 802.11 systems, see e.g. [14], [15], [16]. But again, our problem has one additional dimension (a channel has to be selected): in turn, the number of available decisions at the transmitter is much larger than in 802.11 systems where only the rate has to be chosen. Rate adaptation algorithms are not applicable when the channel can also be selected for each packet transmission. This is due to the fact that the transmitter does not continuously monitor the same channel (as in 802.11 systems), and has to switch channels often to discover the best (channel, rate) pair as rapidly as possible.

There is an abundant literature on MAB problems, and engineers have applied these problems to dynamic spectrum access [8], [10], [11], [17]. Most existing theoretical results, see [18] for a recent survey, are concerned with *unstructured* MAB problems, i.e., problems where the average reward associated with the various decisions are not related. For this kind of problems, Lai and Robbins [6] derived an asymptotic lower bound on regret and also designed optimal sequential decision algorithms. When the average rewards are structured (as this is the case for our problem), the design of optimal decision algorithms is more challenging, see e.g. [18]. Non-stationary environments have not been extensively studied in the bandit literature: Most often unstructured MAB only are analyzed, see [19], [20], [21].

To our knowledge, the only work dealing with joint (channel, rate) selection is [3]. However there, the structural properties of the corresponding MAB problem

had not been identified, and the authors only proposed algorithms based on heuristics. This contrasts with the present work: we rigorously determine fundamental limits satisfied by any (channel, rate) adaptation algorithm, and propose algorithms approaching these limits.

### III. MODELS

We consider a single link (a transmitter-receiver pair). At time 0, the link becomes active and the transmitter starts sending packets to the receiver. For each packet, the transmitter selects a channel from a finite set  $\mathcal{C} = \{1, \dots, C\}$ , and a coding and modulation scheme from a finite set  $\mathcal{R} = \{r_k, k \in \mathcal{K}\}$ , with  $\mathcal{K} = \{1, \dots, K\}$ .  $\mathcal{R}$  is ordered, i.e.,  $r_1 < r_2 < \dots < r_K$ . After a packet is sent, the transmitter is informed of whether the transmission has been successful. Based on the observed past transmission successes and failures at the various channels and rates, the transmitter has to select a channel and rate pair for the next packet transmission. Let  $\Pi$  denote the set of all possible sequential (channel, rate) selection schemes. Packets are assumed to be of equal size, and without loss of generality, for any  $k$ , the duration of a packet transmission at rate  $r_k$  is  $1/r_k$ .

#### A. Channel models

For the  $i$ -th packet transmission on channel  $c$  at rate  $r_k$ , a binary random variable  $X_{ck}(i)$  represents the success ( $X_{ck}(i) = 1$ ) or failure ( $X_{ck}(i) = 0$ ) of the transmission. We consider both stationary and non-stationary radio environments. In stationary environments, the success transmission probabilities on the various channels and at different rates do not evolve over time. This arises when the system considered is static (in particular, the transmitter and receiver do not move). In non-stationary environments, success transmission probabilities can evolve over time. Unless otherwise specified, we consider stationary radio environments. Non-stationary environments are treated in Section VII.

We assume that  $X_{ck}(i)$ ,  $i = 1, 2, \dots$ , are independent and identically distributed, and we denote by  $\theta_{ck}$  the success transmission probability on channel  $c$  at rate  $r_k$ :  $\theta_{ck} = \mathbb{E}[X_{ck}(i)]$ . We verified that the i.i.d. assumption holds in our test-bed and simulation framework. Denote by  $(c^*, k^*)$  the optimal (channel, rate) pair, i.e.,  $(c^*, k^*) \in \arg \max_{c,k} r_k \theta_{ck}$ . To simplify the exposition and the notation, we assume that the optimal (channel, rate) pair is unique, i.e.,  $r_{k^*} \theta_{c^* k^*} > r_k \theta_{ck}$ , for all  $(c, k) \neq (c^*, k^*)$ . We further introduce, for any channel  $c$ , the optimal rate  $r_{k_c^*}$ , i.e.,  $(c, k_c^*) \in \arg \max_k r_k \theta_{ck}$ . Again for simplicity, we assume that on any channel, the optimal rate is unique:  $r_{k_c^*} \theta_{ck_c^*} > r_k \theta_{ck}$ , for all  $k \neq k_c^*$ .

The throughput achieved using (channel, rate) pair  $(c, k)$  is denoted by  $\mu_{ck} = r_k \theta_{ck}$ . The maximum throughput on channel  $c$  is  $\mu_c^* = \mu_{ck_c^*}$ , and the throughput achieved using the optimal (channel, rate) pair is  $\mu^* = \mu_{c^*k^*}^* = \mu_{c^*k^*}$ .

### B. Structural properties

The successful transmission probabilities  $\theta = (\theta_{ck}, c \in \mathcal{C}, k \in \mathcal{K})$  are initially unknown at the transmitter, and have to be learnt. When the number of (channel, rate) pairs grows large, learning the best pair for transmission then becomes really challenging. Fortunately, the outcomes of transmissions using the various (channel, rate) pairs exhibit structural properties that can be exploited to speed up the learning process. To emphasize the importance of exploiting the structural properties, we consider three scenarios with increasing structure.

1) *Scenario 1 – No structure*: If no structural assumptions are made regarding the successful transmission probabilities, then  $\theta \in [0, 1]^{C \times K}$ . In such scenarios, we will show that the performance loss due to the need to explore sub-optimal (channel, rate) pairs scales linearly with the number of channels and rates.

2) *Scenario 2 – Unimodality*: First observe that the successes and failures of transmissions on a given channel at various rates are statistically correlated. Indeed, if a transmission is successful at a high rate, it has to be successful at a lower rate. Similarly, if a low-rate transmission fails, then transmitting at a higher rate would also fail. Formally this means that for any channel  $c$ ,  $\theta_c = (\theta_{c1}, \dots, \theta_{cK}) \in \mathcal{T}$ , where  $\mathcal{T} = \{\eta \in [0, 1]^K : \eta_1 \geq \dots \geq \eta_K\}$ . Then, in practice, it has been observed (and this is confirmed in our numerical experiments) that the throughput achieved on a given channel is a unimodal function of the transmission rate, see e.g. [5], [16]. In other words, for any channel  $c$ ,  $\theta_c \in \mathcal{U}$ , where  $\mathcal{U} = \{\eta \in [0, 1]^K : \exists k_1, r_1 \eta_1 < \dots < r_{k_1} \eta_{k_1}, r_{k_1} \eta_{k_1} > r_{k_1+1} \eta_{k_1+1} > \dots > r_K \eta_K\}$ . In summary in Scenario 2, for any channel  $c$ ,  $\theta_c \in \mathcal{T} \cap \mathcal{U}$ .

3) *Scenario 3 – Graphical unimodality*: We further observe (see Section VIII) that on a given channel, the throughput first grows linearly with the rate (the successful transmission probability is close to 1), and then abruptly decreases to 0. This observation has been made in earlier work, see [14] (the author refers to this scenario as the *steep throughput* scenario), [5]. This knowledge can be exploited to build a relationship between the throughputs achieved on various channels. Indeed, for example, the throughputs observed on two different channels are roughly identical in their growth

phase (when the rates are low and the success probabilities are close to 1). To exploit this observation, we remark that if it holds, the throughput is a *graphically unimodal* function of the (channel, rate) pair as defined below.

We first construct a directed graph  $G = (V, E)$  whose vertices correspond to the (channel, rate) pairs. When  $(d, d') \in E$ , we say that the decision  $d'$  is a neighbor of decision  $d$ , and we define  $\mathcal{N}(d) = \{d' \in V : (d, d') \in E\}$  as the set of neighbors of  $d$ . The throughput or average reward of decision  $d = (c, k)$  is denoted by  $\mu_d = r_k \theta_{ck}$ . Graphical unimodality expresses the fact that when the optimal decision is  $d^* = (c^*, k^*)$ , then for any  $d \in V$ , there exists a path in  $G$  from  $d$  to  $d^*$  along which the expected reward is strictly increasing. In other words there is no *local* maximum in terms of expected reward except at  $d^*$ . The notion of locality is defined through that of neighborhood  $\mathcal{N}(d), d \in V$ . Formally,  $\theta \in \mathcal{U}_G$ , where  $\mathcal{U}_G$  is the set of successful transmission probabilities  $\theta \in [0, 1]^{C \times K}$  such that, if  $d^* = (c^*, k^*) \in \arg \max_{(c,k)} r_k \theta_{ck}$ , for any  $d = (c, k) \in V$ , there exists a path  $(d_0 = d, d_1, \dots, d_p = d^*)$  in  $G$  such that for any  $i = 1, \dots, p$ ,  $\mu_{d_i} > \mu_{d_{i-1}}$ .

Let us now complete the construction of graph  $G$ . The set of edges  $E$  is:  $((c, k), (c, k-1)), ((c, k), (c, k+1))$  and  $((c, k), (c', k)), ((c, k), (c', k+1))$  for all (channel, rate) pair  $(c, k)$ , and all  $c'$ . An example of such a graph  $G$  is presented in Figure 1 – for 2 channels and 4 rates. When the above observation made on  $\theta$  holds (steep scenario as defined in [14]), it is easy to check that the throughput is a graphically unimodal function (w.r.t. graph  $G$ ) of the channel and rate. In all practical cases, beyond the steep throughput scenario, we have actually observed that the graph  $G$  as constructed above had enough edges to guarantee the graphical unimodality of the throughput, see Section VIII.

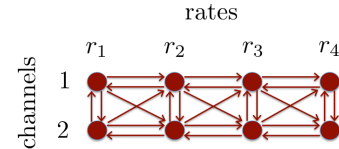


Fig. 1

EXAMPLE OF A GRAPH PROVIDING UNIMODALITY OF THE THROUGHPUT.

In summary, in Scenario 3, we assume that  $\theta \in \mathcal{T}^C \cap \mathcal{U}_G$ . Note that there is more structure in Scenario 3 than in Scenario 2: if  $\theta \in \mathcal{T}^C \cap \mathcal{U}_G$ , then for any  $c$ ,  $\theta_c \in \mathcal{T} \cap \mathcal{U}$ .

#### IV. OBJECTIVES AND MULTI-ARMED BANDITS

Our goal is to devise a sequential (channel, rate) selection scheme that maximizes the number of packets successfully transmitted over a finite time horizon. Such a design can be formulated as an online stochastic optimization problem. The choice of the time horizon, denoted by  $T$ , is not really important as long as during time interval  $T$ , a large number of packets can be sent – so that inferring the success transmission probabilities efficiently is possible.

Consider a rate adaption scheme  $\pi \in \Pi$  that selects (channel, rate) pair  $(c^\pi(t), k^\pi(t))$  for the  $t$ -th packet transmission. The number of packets  $\gamma^\pi(T)$  that have been successfully sent under algorithm  $\pi$  up to time  $T$  is:  $\gamma^\pi(T) = \sum_{c,k} \sum_{i=1}^{s_{ck}^\pi(T)} X_{ck}(i)$ , where  $s_{ck}^\pi(T)$  is the number of transmission attempts on channel  $c$  at rate  $r_k$  before time  $T$ . The  $s_{ck}(T)$ 's are random variables (since the rates selected under  $\pi$  depend on the past random successes and failures), and satisfy the following constraint:

$$\sum_{c,k} s_{ck}^\pi(T) \times \frac{1}{r_k} \leq T.$$

Wald's lemma implies that the expected number of packets successfully sent up to time  $T$  is:  $\mathbb{E}[\gamma^\pi(T)] = \sum_{c,k} \mathbb{E}[s_{ck}^\pi(T)] \theta_{ck}$ . Thus, our objective is to design an algorithm solving the following online stochastic optimization problem:

$$\begin{aligned} \max_{\pi \in \Pi} \quad & \sum_{c,k} \mathbb{E}[s_{ck}^\pi(T)] \theta_{ck}, \\ \text{s.t.} \quad & \sum_{c,k} s_{ck}^\pi(T) \times \frac{1}{r_k} \leq T, \quad \forall c, k, s_{ck}^\pi(T) \in \mathbb{N}. \end{aligned} \quad (1)$$

##### A. An equivalent Multi-Armed Bandit (MAB) problem

Next we show that the above online stochastic optimization problem is equivalent to a Multi-Armed Bandit (MAB) problem.

1) *An alternative system:* Without loss of generality, we assume that time can be divided into slots whose durations are such that for any  $k$ , the time it takes to transmit one packet at rate  $r_k$  corresponds to an integer number of slots. Under this convention, the optimization problem (1) can be written as:

$$\begin{aligned} \max_{\pi \in \Pi} \quad & \sum_{c,k} \mathbb{E}[t_{ck}^\pi(T)] r_k \theta_{ck}, \\ \text{s.t.} \quad & \sum_{c,k} t_{ck}^\pi(T) \leq T, \\ & \forall c, k, t_{ck}^\pi(T) \in \frac{1}{r_k} \mathbb{N} = \{\frac{u}{r_k}, u \in \mathbb{N}\}, \end{aligned} \quad (2)$$

where  $t_{ck}^\pi(T) = s_{ck}^\pi(T)/r_k$  represents the amount of time (in slots) that the transmitter spends, before  $T$ , on sending packets on channel  $c$  at rate  $r_k$ . The constraint  $t_{ck}(T) \in \frac{1}{r_k} \mathbb{N}$  indicates that when a rate is selected, this

rate selection remains the same for the next  $1/r_k$  slots. By relaxing this constraint, we obtain an optimization problem corresponding to a MAB problem. Indeed, consider now an alternative system where rate selection is made *every* slot. If at any given slot, (channel, rate) pair  $(c, k)$  is selected for the  $i$ -th times, then if  $X_{ck}(i) = 1$ , the transmitter successfully sends  $r_k$  bits in this slot, and if  $X_{ck}(i) = 0$ , then no bit are received. A (channel, rate) selection algorithm then decides in each slot which (channel, rate) pair to use. There is a natural mapping between rate selection algorithms in the original system and in the alternative system: let  $\pi \in \Pi$ , if for the  $t$ -th packet transmission, rate  $r_k$  is selected under  $\pi$  in the original system, then  $\pi$  selects the same rate  $r_k$  in the  $t$ -th slot.

For the alternative system, the objective is to design  $\pi \in \Pi$  solving the following optimization problem, which can be interpreted as a relaxation of (2).

$$\begin{aligned} \max_{\pi \in \Pi} \quad & \sum_{c,k} \mathbb{E}[t_{ck}^\pi(T)] r_k \theta_{ck}, \\ \text{s.t.} \quad & \sum_{c,k} t_{ck}^\pi(T) \leq T, \\ & \forall c, k, t_{ck}^\pi(T) \in \mathbb{N}. \end{aligned} \quad (3)$$

The above optimization problem corresponds to a MAB problem, where in each slot a decision is taken (i.e., a channel and a rate are selected), and where when  $(c, k)$  is chosen, the obtained reward is  $r_k$  with probability  $\theta_{ck}$  and 0 with probability  $1 - \theta_{ck}$ .

2) *Regrets:* We quantify the performance of an algorithm  $\pi \in \Pi$  in both original and alternative systems through the notion of *regret*. The regret up to slot  $T$  compares the performance of  $\pi$  to that achieved by an algorithm always selecting the best (channel, rate) pair. If the parameter  $\theta = (\theta_{ck}, c, k)$  was known, then in both systems, it would be optimal to always select (channel, rate) pair  $(c^*, k^*)$ . The regret of algorithm  $\pi$  up to time slot  $T$  in the original system is then defined by:

$$R_1^\pi(T) = \theta_{c^*k^*} \lfloor r_{k^*} T \rfloor - \sum_{c,k} \theta_{ck} \mathbb{E}[s_{ck}^\pi(T)],$$

where  $\lfloor x \rfloor$  denotes the largest integer smaller than  $x$ .

The regret of algorithm  $\pi$  up to time slot  $T$  in the alternative system is similarly defined by:

$$R^\pi(T) = \theta_{c^*k^*} r_{k^*} T - \sum_{c,k} \theta_{ck} r_k \mathbb{E}[t_{ck}^\pi(T)].$$

3) *Asymptotic equivalence:* In the next section, we show that an asymptotic lower bound for the regret  $R^\pi(T)$  is of the form  $c(\theta) \log(T)$  where  $c(\theta)$  is a strictly positive constant that we can explicitly characterize. It means that for all  $\pi \in \Pi$ ,  $\liminf_{T \rightarrow \infty} R^\pi(T)/\log(T) \geq c(\theta)$ . It can be also shown that there exists an al-

gorithm  $\pi^* \in \Pi$  that actually achieves this lower bound in the alternative system, in the sense that  $\limsup_{T \rightarrow \infty} R^{\pi^*}(T)/\log(T) \leq c(\theta)$ . In such a case, we say that  $\pi^*$  is asymptotically optimal. The following proposition states that actually, the same lower bound is valid in the original system, and that any asymptotically optimal algorithm in the alternative system is also asymptotically optimal in the original system.

*Proposition 1:* Let  $\pi \in \Pi$ . For any  $\beta > 0$ , we have:

$$\liminf_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} \geq \beta \implies \liminf_{T \rightarrow \infty} \frac{R_1^\pi(T)}{\log(T)} \geq \beta,$$

and

$$\limsup_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} \leq \beta \implies \limsup_{T \rightarrow \infty} \frac{R_1^\pi(T)}{\log(T)} \leq \beta.$$

**Proof.** Let  $T > 0$ . By time  $T$ , we know that there have been at least  $\lfloor Tr_1 \rfloor$  transmissions, but no more than  $\lceil Tr_K \rceil$ . Also observe that both regrets  $R^\pi$  and  $R_1^\pi$  are increasing functions of time. We deduce that:

$$R^\pi(\lfloor Tr_1 \rfloor) \leq R_1^\pi(T) \leq R^\pi(\lceil Tr_K \rceil).$$

Now

$$\begin{aligned} \liminf_{T \rightarrow \infty} \frac{R_1^\pi(T)}{\log(T)} &\geq \liminf_{T \rightarrow \infty} \frac{R^\pi(\lfloor Tr_1 \rfloor)}{\log(T)} \\ &= \liminf_{T \rightarrow \infty} \frac{R^\pi(\lfloor Tr_1 \rfloor)}{\log(\lfloor Tr_1 \rfloor)} \geq \beta. \end{aligned}$$

The second statement can be derived similarly.  $\square$

### B. MAB problems

Instead of trying to solve (1), we rather focus on analyzing the MAB problem (3). We know that optimal algorithms for (3) will also be optimal for the original problem. The nature of the MAB problem (3) depends on the structural assumption made on the successful transmission probabilities  $\theta$ . In absence of such assumption (Scenario 1), we get a classical MAB problem where the rewards provided by all decisions are independent. In Scenarios 2 and 3, we get a *structured* MAB problem, as we know a priori that  $\theta$  belongs to a structured set, which helps learning the best (channel, rate) pair. Next we summarize the MAB problems obtained in the three different scenarios.

We have a set  $\{1, \dots, C\} \times \{1, \dots, K\}$  of possible decisions (i.e., (channel, rate) pairs). If decision  $(c, k)$  is taken for the  $i$ -th time, we receive a reward  $r_k X_{ck}(i)$ . ( $X_{ck}(i), i = 1, 2, \dots$ ) are i.i.d. with Bernoulli distribution with mean  $\theta_{ck}$ . The objective is to design a decision scheme minimizing the regret  $R^\pi(T)$  over all possible algorithms  $\pi \in \Pi$ . The three MAB problems differ depending on the structural assumptions made on  $\theta$ .

**Unstructured MAB** ( $P_I$ ). No assumption is made on  $\theta$ :  $\theta \in [0, 1]^{C \times K}$ .

**Structured MAB** ( $P_U$ ). We assume that  $\theta_c \in \mathcal{T} \cap \mathcal{U}$  for all channel  $c$ .

**Structured MAB** ( $P_{GU}$ ). We assume that  $\theta \in \mathcal{T}^C \cap \mathcal{U}_G$ .

## V. REGRET LOWER BOUNDS

In this section, we derive an asymptotic (as  $T$  grows large) lower bound of the regret  $R^\pi(T)$  satisfied by any algorithm  $\pi \in \Pi$  in the three MAB bandit problems ( $P_I$ ), ( $P_U$ ), and ( $P_{GU}$ ). These lower bounds provide insightful theoretical performance limits satisfied by any (channel, rate) selection scheme. By comparing the lower bounds derived for the three problems, we also quantify the performance gains that can be achieved by smartly exploiting the (a priori) known structure.

### A. Unstructured MAB ( $P_I$ )

The regret lower bound for MAB problem ( $P_I$ ) can be derived using the direct technique used by Lai and Robbins [6]. Note that the only difference between ( $P_I$ ) and the classical MAB problems [6] lies in the fact that in ( $P_I$ ), we know that the average reward of decision  $(c, k)$  is of the form  $r_k \theta_{ck}$  for known  $r_k$ . The analysis of ( $P_I$ ) is then similar to that of classical bandit problems.

We first introduce the notion of uniformly good algorithms. An algorithm  $\pi$  is uniformly good, if for all parameters  $\theta$ , for any  $\alpha > 0$ , we have<sup>1</sup>:  $\mathbb{E}[t_{ck}^\pi(T)] = o(T^\alpha), \forall (c, k) \neq (c^*, k^*)$ , where  $t_{ck}^\pi(T)$  is the number of times (channel, rate) pair  $(c, k)$  has been chosen up to the  $T$ -th decision, and  $(c^*, k^*)$  is the optimal channel and rate pair (it depends on  $\theta$ ). Uniformly good algorithms exist as we shall see later on.

Let  $N = \{k : \mu^* \leq r_k\}$  – note that  $N$  depends on  $\theta$ . There exists  $k_0$  such that  $N = \{k_0, \dots, K\}$ , with the convention  $k_0 = K + 1$  if  $N = \emptyset$ . Note that if  $k < k_0$ , then for any channel  $c$ ,  $r_{k^*} \theta_{ck^*} > r_k$ , which means that even if all transmissions at rate  $r_k$  on channel  $c$  were successful, i.e.,  $\theta_{ck} = 1$ , rate  $r_k$  would be sub-optimal. Hence, there is no need to select rate  $r_k$  to discover this fact, since by only selecting rate  $r_{k^*}$  on channel  $c^*$ , we get to know whether  $r_{k^*} \theta_{c^* k^*} > r_k \geq r_k \theta_k$ .

Finally, we introduce the Kullback-Leibler (KL) divergence, a well-known measure for dissimilarity between two distributions. When we compare two Bernoulli distributions with respective averages  $p$  and  $q$ , the KL divergence is:  $I(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q}$ .

<sup>1</sup>  $f(T) = o(g(T))$  means that  $\lim_{T \rightarrow \infty} f(T)/g(T) = 0$ .

*Theorem 1:* Let  $\pi \in \Pi$  be a uniformly good rate selection algorithm for MAB problem ( $P_I$ ). We have:  $\liminf_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} \geq c_I(\theta)$ , where

$$c_I(\theta) = \sum_{k=k_0: k \neq k^*}^K \frac{\mu^* - r_k \theta_{c^*k}}{I(\theta_{c^*k}, \frac{\mu^*}{r_k})} + \sum_{c \neq c^*} \sum_{k=k_0}^K \frac{\mu^* - r_k \theta_{ck}}{I(\theta_{ck}, \frac{\mu^*}{r_k})}.$$

The proof of the previous theorem is similar to that of the regret lower bound in [6], and is omitted here. In view of this result, if we do not exploit structural properties of the problem, then the regret of any algorithm scales at least as  $CK \log(T)$ . Hence, when the number of channels and rates grow large, no algorithm is able to learn the best (channel, rate) pair rapidly and efficiently.

### B. Structured MAB ( $P_U$ )

To derive a regret lower bound for MAB problem ( $P_U$ ), we need to introduce additional notations. We define  $M = N \cap \{k^* - 1, k^* + 1\}$ . For any channel  $c$ , let  $N_c = \{k : \mu_c^* \leq r_k\}$ , and  $k_{0c}$  such that  $N_c = \{k_{0c}, \dots, K\}$ , with the convention  $k_{0c} = K + 1$  if  $N_c = \emptyset$ . Observe that for any  $c \neq c^*$ ,  $k_{0c} \leq k_0$ . Let  $M_c = N_c \cap \{k_c^* - 1, k_c^* + 1\}$ .

*Theorem 2:* Let  $\pi \in \Pi$  be a uniformly good (channel, rate) selection algorithm for MAB problem ( $P_U$ ). We have:  $\liminf_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} \geq c_U(\theta)$ , where  $c_U(\theta)$  is the optimal value of the following optimization problem:

$$\begin{aligned} & \inf_{\alpha_{ck} \geq 0, \forall k, c} \sum_{(c,k) \neq (c^*, k^*)} \alpha_{ck} (\mu^* - \mu_{ck}) \\ \text{s.t. } & \forall k \in M, \alpha_{c^*k} I(\theta_{c^*k}, \frac{\mu^*}{r_k}) \geq 1, \\ & \forall c \neq c^* : k_c^* \geq k_{0c}, \alpha_{ck_c^*} I(\theta_{ck_c^*}, \frac{\mu^*}{r_{k_c^*}}) \geq 1, \text{ and} \\ & \forall k \geq k_0, k \neq k^*, \inf_{\lambda_c \in C_k} \sum_l \alpha_{cl} I(\theta_{cl}, \lambda_{cl}) \geq 1, \end{aligned}$$

where  $C_k = \{\lambda_c \in \mathcal{U} \cap \mathcal{T} : r_k \lambda_{ck} > \mu^*\}$ .

All proofs are presented in Appendix. The above theorem does not provide a fully explicit regret lower bound. In particular, it remains unclear how this lower bound scales with the numbers of rates and channels. In the following theorem, we further exploit the structural properties of the MAB problem ( $P_U$ ) to show that  $c_U(\theta)$  scales at most linearly with the number of channels, and does not scale with the number of rates.

*Theorem 3:* We have  $c_U(\theta) \leq c'_U(\theta)$  where

$$\begin{aligned} c'_U(\theta) = & \sum_{k \in M} \frac{\mu^* - \mu_{c^*k}}{I(\theta_{c^*k}, \frac{\mu^*}{r_k})} \\ & + \sum_{c \neq c^*} \left[ \frac{\mu^* - \mu_{ck_c^*}}{\min\{I(\theta_{ck_c^*}, \frac{\mu^*}{r_{k_c^*}}), I(\theta_{ck_c^*}, \theta_{ck_c^*} - \frac{\delta_c}{r_{k_c^*}})\}} \right. \\ & \left. + \sum_{k \in M_c} \frac{\mu^* - \mu_{ck}}{I(\theta_{ck}, \theta_{ck} + \frac{\delta_c}{r_k})} \right]. \end{aligned} \quad (4)$$

and

$$\delta_c = \min_{k \in \{k_c^* - 1, k_c^* + 1\}} (\mu_{ck_c^*} - \mu_{ck}) / 2.$$

In particular,  $c'_U(\theta)$  is proportional to the number of channels and independent of the number of rates.

From the above analysis, we conclude that the minimum regret for the MAB problem ( $P_U$ ) scales at most as  $3C \log(T)$ . Hence we expect that exploiting the structure of the problem (the fact that  $\theta_c \in \mathcal{T} \cap \mathcal{U}$  for any channel  $c$ ) may significantly improve the system performance. Indeed we expect a regret that does not depend on the number of available rates. In the next section, we design an algorithm with such a regret.

### C. Structured MAB problem ( $P_{GU}$ )

Graphical unimodal bandit problems have been recently studied in [22], [23]. A regret lower bound is derived in [23]. The only difference between our graphically unimodal MAB problem and those considered in [23] is that we consider directed graphs, but the analysis is similar. We use here the notation introduced in §III-B.3, and recall that  $N = \{k : \mu^* \leq r_k\}$ . For any  $(c, k)$ , we define  $\mathcal{N}'(c, k) = \mathcal{N}(c, k) \cap N$ .  $\mathcal{N}'(c, k)$  is the set of (channel, rate) pairs that are neighbors of vertex  $(c, k)$ , and that need to be explored if one wants to know whether they provide better throughput than  $(c, k)$ .

*Theorem 4:* [23] Let  $\pi \in \Pi$  be a uniformly good (channel, rate) selection algorithm for MAB problem ( $P_{GU}$ ). We have:

$$\liminf_{T \rightarrow +\infty} \frac{R^\pi(T)}{\log(T)} \geq c_{GU}(\theta), \quad (5)$$

where

$$c_{GU}(\theta) = \sum_{(c,k) \in \mathcal{N}'(c^*, k^*)} \frac{\mu^* - \mu_{ck}}{I(\theta_{ck}, \frac{\mu^*}{r_k})}.$$

In view of the above theorem, for the MAB problem ( $P_{GU}$ ), the minimum regret scales as  $\gamma \log(T)$ , where  $\gamma$  is the maximum node degree in the graph  $G$ . Note that for our graph  $G$ ,  $\gamma \leq 2C$ . Hence, by exploiting the graphical unimodal structure, we may expect to design algorithms whose regret does not depend on the number

of available rates. In the next section, an algorithm whose regret matches the lower bound of Theorem 4 is proposed.

In this section, we have shown that the regret lower bound can be significantly improved when structural assumptions are made, i.e.,  $c_{GU}(\theta) \leq c_U(\theta) \leq c_I(\theta)$ . By exploiting the structure of the problem, we may actually design algorithms whose regrets does not depend on the number of available rates. Such algorithms do not exist when the structure is not exploited (see Theorem 1).

## VI. ALGORITHMS

In this section, we present algorithms for the three MAB problems  $(P_I)$ ,  $(P_U)$ , and  $(P_{GU})$ , and analyze their regrets. For the two structured MAB problems, the proposed algorithms exhibit a regret that does not depend on the number of available rates.

### A. The KL-UCB algorithm for MAB problem $(P_I)$

Classical unstructured bandit problems have been extensively studied in the past, and numerous efficient algorithms have been proposed. We build on this previous work, and present a simple extension of KL-UCB algorithm [21] to the MAB problem  $(P_I)$ . This algorithm does not exploit any structural properties, and is asymptotically optimal: its regret matches the lower bound derived in Theorem 1.

Under the KL-UCB algorithm, each (channel, rate) pair  $(c, k)$  is associated with an index  $q_{ck}(n)$  for the  $(n+1)$ -th packet transmission:

$$q_{ck}(n) = \max\{q \in [0, r_k] : t_{ck}(n)I\left(\frac{\hat{\mu}_{ck}(n)}{r_k}, \frac{q}{r_k}\right) \leq \log(n) + 3\log\log(n)\},$$

where  $t_{ck}(n)$  denotes the number of times  $(c, k)$  has been selected up to the  $n$ -th transmission, and

$$\hat{\mu}_{ck}(n) = \frac{1}{t_{ck}(n)} \sum_{i=1}^{t_{ck}(n)} r_k X_{ck}(i),$$

is the empirical throughput or reward of (channel, rate) pair  $(c, k)$  up to the  $n$ -th transmission. The algorithm selects the (channel, rate) pair with highest index:

---

#### Algorithm 1 KL-UCB

---

For  $n = 0, \dots, CK - 1$  (initialization): for the  $(n+1)$ -th transmission, select (channel, rate) pair  $(c, k)(n+1) = (c' + 1, k' + 1)$  where  $n = Kc' + k'$ ,  $k' \in \{0, \dots, K-1\}$ . For  $n \geq CK$ , for the  $(n+1)$ -th transmission, select  $(c, k)(n+1)$  where  $(c, k)(n+1) \in \arg \max_{(c,k)} q_{ck}(n)$ .

---

KL-UCB is known to be asymptotically optimal in classical bandit problems [21]. It can be easily established that its extension is also optimal for the problem  $(P_I)$ :

*Theorem 5:* For any  $\theta \in [0, 1]^{C \times K}$ , the regret of the  $\pi = \text{KL-UCB}$  algorithm satisfies:

$$\limsup_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} \leq c_I(\theta),$$

In particular, the regret under KL-UCB scales linearly with the numbers of channels and rates. When the later become large, the performance of KL-UCB can be quite poor.

### B. The CRS-T algorithm for MAB problem $(P_U)$

Next, we present CRS-T (Channel and Rate Sampling with Tests), an algorithm that exploits the structure of the MAB problem  $(P_U)$ , i.e., the fact that on each channel, the throughput is a unimodal function of the rate. To describe our algorithm, we introduce the following notations. After the  $n$ -th transmission, the rate with the highest average empirical throughput on channel  $c$  is referred to as the *leader* on channel  $c$ , and is  $l_c(n) = \arg \max_k \hat{\mu}_{ck}(n)$ . The *global leader*  $l(n)$  is the (channel, rate) pair with highest average empirical throughput:  $l(n) = \arg \max_{(c,k)} \hat{\mu}_{ck}(n)$ .

We also introduce the following statistical tests, which will be used to assess whether the leader on channel  $c$ ,  $l_c(n)$ , provides a larger reward than its neighbors  $l_c(n) - 1, l_c(n) + 1$  on the same channel. Let  $N(k) = \{k-1, k+1\}$  denote the set of neighbors of rate  $r_k$ , and define:

$$\delta_{ck}(n) = \hat{\mu}_{cl_c(n)}(n) - \hat{\mu}_{ck}(n),$$

$$\bar{\delta}_{ck}(n) = \sqrt{\alpha \log(T) \left( \frac{r_k^2}{2t_{ck}(n)} + \frac{r_{l_c(n)}^2}{2t_{cl_c(n)}(n)} \right)}.$$

for some constant  $\alpha > 1$ . The test for channel  $c$  at time  $n$  is defined through  $U_c(n) \in \{-1, 0, 1\}$ :

$$U_c(n) = \begin{cases} 1 & \text{if } \max_{k \in N(l_c(n))} (\delta_{ck}(n) - \bar{\delta}_{ck}(n)) > 0, \\ -1 & \text{if } \min_{k \in N(l_c(n))} (\delta_{ck}(n) + \bar{\delta}_{ck}(n)) < 0, \\ 0 & \text{otherwise.} \end{cases}$$

The test has to be interpreted as follows.  $U_c(n) = 1$  means that  $l_c(n)$  is better than its neighbors with high probability,  $U_c(n) = -1$  means that  $l_c(n)$  has at least one neighbor which is better with high probability, and  $U_c(n) = 0$  means that we do not have enough samples to determine whether  $l_c(n)$  is better than its neighbors. After the  $n$ -th packet transmission, we define  $\mathcal{U}(n) = \{c : U_c(n) = 0\}$  the set of channels for which we



cannot determine whether the leader  $l_c(n)$  corresponds to the best rate  $k_c^*$  on this channel. We also define the set  $\mathcal{V}_c(n) = \{k \in N(l_c(n)), \delta_{ck}(n) - \bar{\delta}_{ck}(n) < 0\}$ , the set of neighboring rates which might still be better than the leader  $l_c(n)$  on channel  $c$ .

The sequential decisions under the CRS-T algorithm are based on the indexes of the various (channel, rate) pairs, and can be easily implemented. The index  $b_k(n)$  of decision  $(c, k)$  for the  $(n+1)$ -th packet transmission is:

$$b_{ck}(n) = q_{ck}(n) \mathbf{1}\{(k = l_c(n)) \cup (U_c(n) \neq 1)\},$$

where  $q_{ck}(n)$  is the index used in the KL-UCB algorithm. Note that the index of decision  $(c, k)$  is equal to 0 if  $k$  is not the leader on channel  $c$  and if  $U_c(n) \neq 1$ . The pseudo-code for CRS-T is given below (each time the decision is ambiguous, ties are broken arbitrarily).

---

**Algorithm 2** CRS-T

---

For  $n = 0, \dots, CK - 1$  (initialization): for the  $(n+1)$ -th transmission, select (channel, rate) pair  $(c, k)(n+1) = (c' + 1, k' + 1)$  where  $n = Kc' + k'$ ,  $k' \in \{0, \dots, K-1\}$ . For  $n \geq CK$ : for the  $(n+1)$ -th transmission, select  $(c, k)(n+1)$  where

- if  $\min_{c,k} t_{ck}(n) < \log(\log(n))$ ,  
 $(c, k)(n+1) \in \arg \min_{c',k'} t_{c'k'}(n)$ ;
  - else
    - if  $\mathcal{U}(n) \neq \emptyset$ , then  $c(n+1) \in \mathcal{U}(n)$  and  
 $k(n+1) \in \arg \min_{k' \in \mathcal{V}_{c(n)}(n)} t_{c(n)k'}(n)$ ;
    - else  $(c, k)(n+1) \in \arg \max_{c',k'} b_{c'k'}(n)$ .
- 

The design of the CRS-T algorithm is motivated by the following objectives: (1) We explore each decision at least  $\log \log(n)$  times before the  $n$ -th transmission. This makes sure that all decisions are selected infinitely many times so that the empirical averages  $\hat{\mu}_{ck}(n)$  converge a.s. to their true value  $\mu_{ck}$  when  $n \rightarrow +\infty$ . (2) For all channels, we need to play the leader and all its neighbours until we can determine with high probability whether the leader  $l_c(n)$  is the best rate  $k_c^*$ . (3) If for a given channel  $c$ , we have determined that  $l_c(n)$  is  $k_c^*$ , then we play only  $l_c(n)$  on channel  $c$ . (4) If for a given channel  $c$ , we have determined that  $l_c(n)$  is not  $k_c^*$ , then we play all rates on channel  $c$ , ignoring the unimodal structure.

The next theorem provides an asymptotic upper bound on the regret under CRS-T. This bound scales linearly with the number of channels, but is independent of the number of available rates. In particular, CRS-T efficiently exploits the structure of the MAB problem  $(P_U)$ .

*Theorem 6:* For any  $\theta$  such that for all  $c$ ,  $\theta_c \in \mathcal{T} \cap \mathcal{U}$ , the regret of the CRS-T algorithm satisfies:

$$\limsup_{T \rightarrow \infty} \frac{R^{\text{CRS-T}}(T)}{\log(T)} \leq c^{\text{CRS-T}}(\theta),$$

with

$$c^{\text{CRS-T}}(\theta) = \sum_c \sum_{k \in N(k_c^*)} (\mu^* - \mu_{ck}) \tau_{ck} + \sum_{c \neq c^*} (\mu^* - \mu_{ck_c^*}) \max \left( \frac{1}{I(\theta_{ck_c^*}, \mu^*/r_{k_c^*})}, \max_{k \in N(k_c^*)} \tau_{ck} \right),$$

and

$$\tau_{ck} = \frac{\alpha(r_k^2 + r_{k_c^*}^2)}{2(\mu_{ck_c^*} - \mu_{ck})^2}.$$

The regret under CRS-T does not depend on the number of available rates, and hence efficiently exploits (at least asymptotically) the unimodal structure of  $(P_U)$ .

### C. The KL-UCB-U algorithm for MAB problem $(P_{GU})$

Finally, we present KL-UCB-U, an algorithm for MAB problem  $(P_{GU})$ . KL-UCB-U is a natural extension of an algorithm proposed in [23] for graphically unimodal bandits with undirected graphs. This algorithm is asymptotically optimal (its regret matches the lower bound derived in Theorem 4).

Recall that the global leader is denoted by  $l(n)$  before the  $(n+1)$ -th transmission. We introduce  $v_{(c,k)}(n)$  the number of times that (channel, rate) pair  $(c, k)$  has been the global leader up to the  $n$ -th transmission:  $v_{(c,k)}(n) = \sum_{n'=1}^n \mathbf{1}\{l(n') = (c, k)\}$ . The index associated with decision  $(c, k)$  before the  $(n+1)$ -th transmission is:

$$b_{ck}(n) = \max \left\{ q \in [0, r_k] : t_{ck}(n) I\left(\frac{\hat{\mu}_{ck}(n)}{r_k}, \frac{q}{r_k}\right) \leq \log(v_{l(n)}(n)) + 3 \log(\log(v_{l(n)}(n))) \right\},$$

For the  $(n+1)$ -th transmission, KL-UCB-U selects the (channel, rate) pair in the neighborhood of the leader with maximum index. Ties are broken arbitrarily.

---

**Algorithm 3** KL-UCB-U

---

For  $n = 0, \dots, CK - 1$  (initialization): for the  $(n+1)$ -th transmission, select (channel, rate) pair  $(c, k)(n+1) = (c' + 1, k' + 1)$  where  $n = Kc' + k'$ ,  $k' \in \{0, \dots, K-1\}$ . For  $n \geq CK$ : for the  $(n+1)$ -th transmission, select  $(c, k)(n+1)$  where:

$$(c, k)(n+1) = \begin{cases} l(n) & \text{if } (v_{l(n)}(n) - 1)/\gamma \in \mathbb{N}, \\ \arg \max_{(c,k) \in \mathcal{N}(l(n))} b_{ck}(n) & \text{otherwise.} \end{cases}$$


---

Remember that  $\gamma$  is the maximum number neighbors in  $G$  of a given (channel, rate) pair. The KL-UCB-U algorithm periodically selects the leader to make sure that the latter is often selected. As in [23], we can establish that KL-UCB-U is asymptotically optimal:

*Theorem 7:* For any  $\theta \in \mathcal{T}^C \cap \mathcal{U}_G$ , the regret of  $\pi = \text{KL-UCB-U}$  satisfies:

$$\limsup_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} \leq c_{GU}(\theta),$$

In particular, KL-UCB-U optimally exploits the structure of MAB problem ( $P_{GU}$ ). In turn, if the throughput is a graphically unimodal function of the (channel, rate) pair, then KL-UCB-U asymptotically outperforms any other algorithm, and in particular CRS-T, an algorithm designed to exploit the unimodal structure per channel only.

## VII. NON-STATIONARY RADIO ENVIRONMENTS

In practice, channel conditions may be non-stationary, i.e., the success probabilities at various (channel, rate) pair could evolve over time. In many situations, the evolution over time is rather slow – refer to [3] and to Section V for test-bed measurements. These slow variations allow us to devise (channel,rate) adaptation schemes that efficiently track the best (channel,rate) pair for transmission.

We assume that for all  $(c, k)$  pairs, the transmissions outcomes  $X_{ck}(n)$ ,  $n = 1, 2, \dots$  are independent, with expectation  $\theta_{ck}(n) = \mathbb{E}[X_{ck}(n)]$ . At time  $n$  we define the throughput of  $(c, k)$   $\mu_{ck}(n) = r_k \theta_{ck}(n)$ , the best throughput  $\mu^*(n) = \max_{c,k} \mu_{ck}(n)$  and the optimal decision  $(c^*, r^*)(n) = \arg \max_{c,k} \mu_{ck}(n)$ .

Any algorithm designed for stationary radio environments can readily be extended to non-stationary environments. These extensions are obtained by replacing empirical averages by averages over a sliding time window. Let  $\tau \geq 1$  denote the sliding window size, and define the empirical reward  $\hat{\mu}_{ck}^\tau(n)$  as:

$$\hat{\mu}_{ck}^\tau(n) = \frac{r_k}{t_{ck}^\tau(n)} \sum_{n'=n-\tau+1}^n X_{ck}(n') \mathbf{1}\{(c, k)(n') = (c, k)\},$$

where

$$t_{ck}^\tau(n) = \sum_{n'=n-\tau+1}^n \mathbf{1}\{(c, k)(n') = (c, k)\},$$

with the convention  $\hat{\mu}_{ck}^\tau(n) = 0$  if  $t_{ck}^\tau(n) = 0$ . We also define the upper confidence index of (channel, rate) pair

$(c, k)$  as:

$$q_{ck}^\tau(n) = \sup\{q \in [\hat{\mu}_{ck}^\tau(n), r_k] : I(\frac{\hat{\mu}_{ck}^\tau(n)}{r_k}, \frac{q}{r_k}) \leq \log(\tau) + 3 \log(\log(\tau))\}.$$

We define sliding window variants of the algorithms presented in Section VI by replacing  $t_{ck}(n)$  by  $t_{ck}^\tau(n)$ ,  $\hat{\mu}_{ck}(n)$  by  $\hat{\mu}_{ck}^\tau(n)$  and  $q_{ck}(n)$  by  $q_{ck}^\tau(n)$ . For instance, KL-UCB with sliding window is the algorithm which selects  $(c, k)(n) \in \arg \max_{c,k} q_{ck}^\tau(n)$ .

In [24], the authors show that algorithms with sliding windows efficiently track the best decision over time provided that the environment evolves relatively slowly. This is confirmed in [23], where the performance of algorithms similar to KL-UCB and KL-UCB-U with sliding window is analyzed. Due to space limitation, we skip this analysis; refer to [23] for more details.

## VIII. NUMERICAL EXPERIMENTS

In this section we numerically illustrate the performance of the proposed (channel, rate) selection algorithms. We provide both trace-driven results, where traces are extracted from a real test-bed [3], and simulation results based on a model for the propagation of radio waves and a mapping between channel quality and probability of packet successful transmission on a given (channel,rate) pair [5].

### A. Test-bed experiments

In this subsection we present trace-driven experiments using the test-bed described in [3]. The test-bed is based on a SDR platform (Lyrtech SFF-SDR), and is located in an indoor office. The PHY layer is OFDM, as in 802.11a/g/n. There are 3 available rates  $\{4.5, 6, 6.75\}$  Mbps corresponding to QPSK modulation with respective coding rates  $\{1/2, 2/3, 3/4\}$ . We consider 5 channels in the UHF band centred at  $\{510, 530, 550, 580, 600\}$  Mhz. The bandwidth of each channel is 10 Mhz, and the packet size is 1500 bytes. The trace duration is 600s.

In Fig.2, we plot the best decision  $(c^*, k^*)$  as a function of time. the radio environment is non-stationary, and the optimal decision remains constant for several seconds. Since a packet transmission lasts about 1ms, the packet successful transmission probabilities for various decisions stay constant for thousands of packet transmissions. Therefore we have quite a lot of statistical information to find the best decision. Furthermore the window size used in the tested algorithms should be of the order of a few seconds – we fix it to 2s.

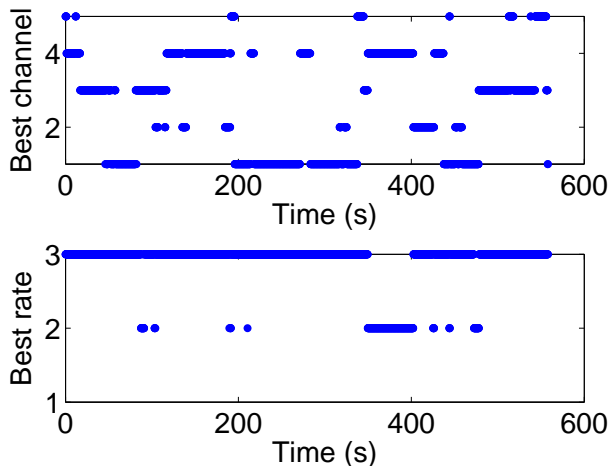


Fig. 2

TEST-BED: BEST (CHANNEL, RATE) PAIR  $(c^*, k^*)(t)$  AS A FUNCTION OF TIME.

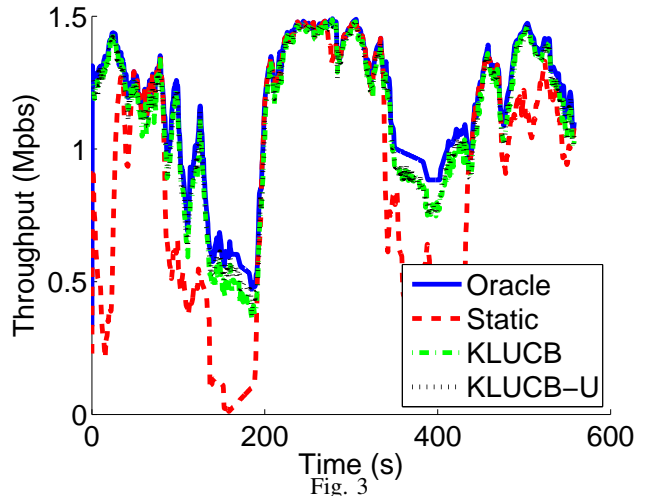


Fig. 3

TEST-BED: THROUGHPUTS OF THE VARIOUS ALGORITHMS AS A FUNCTION OF TIME.

In Fig.3, we plot the throughput under KL-UCB and KL-UCB-U algorithms. For the sake of comparison, we also plot  $\mu^*(t)$  the throughput of an Oracle algorithm that always selects the optimal decision. We also plot the throughput obtained by choosing the best static (channel, rate) pair, computed offline. We observe that selecting the best static pair is clearly sub-optimal, so that adaptive algorithms can lead to a large gain in throughput. Both decision algorithms, KL-UCB and KL-UCB-U, manage to closely follow the best (channel, rate) pair. KL-UCB-U provides a throughput equal to 95% of that obtained under the Oracle algorithm, whereas the throughput under KL-UCB is equal to 90% of that of the Oracle algorithm. There is not a huge performance gap between KL-UCB and KL-UCB-U because there are few available rates,  $K = 3$ . Hence KL-UCB explores  $C \times K = 15$  (channel,rate) pairs, while KL-UCB-U explores (in the worse case)  $2C + 1 = 11$  pairs. We will show that increasing the number of available rates  $K$  makes this difference significantly larger.

### B. Simulation-based experiments

We also present numerical results based on a widely used statistical model for radio propagation. Namely, we assume that the channel is a multi-path Rayleigh fading channel. When a signal is transmitted, several delayed copies of this signal are received and the amplitude and phase of each delayed copy is an independent Rayleigh fading process. We use Jakes' model to simulate Rayleigh fading with user speed set to match the time variability of the test-bed trace presented in VIII-A. This corresponds to static users such as laptops in an

office environment. The expected power of each delayed path is chosen according to the field measurements presented in [25].

We assume that OFDM is used, and the mapping between the strength of received signal on each sub-carrier and the probability of successful transmission is calculated by the method presented in [5]. We consider 5 channels with bandwidth 20 Mhz in the 2.4 GHz band centred at  $\{2.4, 2.41, 2.42, 2.43, 2.44\}$  GHz, respectively. Each channel has 52 sub-carriers and the packet size is 1500 bytes. We consider 8 available rates:  $\{6, 13, 19.5, 26, 39, 52, 58.5, 65\}$  Mbps, and a transmitter-receiver pair with an average SNR of 20 dB. The trace length is 600 seconds.

We first consider stationary environments, so that a snapshot of the success probabilities for all (channel,rate) pairs is drawn and kept constant throughout the simulation. Fig.4 shows the packet successful transmission probabilities and throughputs of different (channel,rate) pairs. As announced, graphical unimodality holds: the throughput on each channel is a unimodal function of the rate, and given the optimal rate  $k_c^*$  on sub-optimal channel  $c \neq c^*$ , there exists another channel  $c' \neq c$  such that either  $\mu_{c',k_c^*} > \mu_{c,k_c^*}$  or  $\mu_{c',k_c^*+1} > \mu_{c,k_c^*}$ . Graphical unimodality results from the fact that we are in a steep environment as defined in [14]. Fig.5 presents the regret of KL-UCB and KL-UCB-U as a function of time. KL-UCB-U beats KL-UCB and for large time horizons the regret under KL-UCB-U is roughly half of that under KL-UCB. Hence exploiting the graphical unimodal structure significantly helps.

We now turn to non-stationary environments. As

$r_k$	6	13	19.5	26	39	52	58.5	65
$\theta_{1,k}$	1	1	1	1	1	0.2	0	0
$\theta_{2,k}$	1	1	1	1	1	1	0.7	0.1
$\theta_{3,k}$	1	1	1	1	1	0.6	0	0
$\theta_{4,k}$	0	0	0	0	0	0	0	0
$\theta_{5,k}$	1	1	0.8	0.2	0	0	0	0
$\mu_{1,k}$	6	13	19.5	26	39	13	0	0
$\mu_{2,k}$	6	13	19.5	26	39	52	41	8
$\mu_{3,k}$	6	13	19.5	26	39	29	0	0
$\mu_{4,k}$	0	0	0	0	0	0	0	0
$\mu_{5,k}$	6	13	16	6	0	0	0	0

Fig. 4

SIMULATION: PACKET SUCCESSFUL TRANSMISSION PROBABILITIES AND THROUGHPUTS AT DIFFERENT (CHANNEL,RATE) PAIRS IN A STATIONARY ENVIRONMENT.

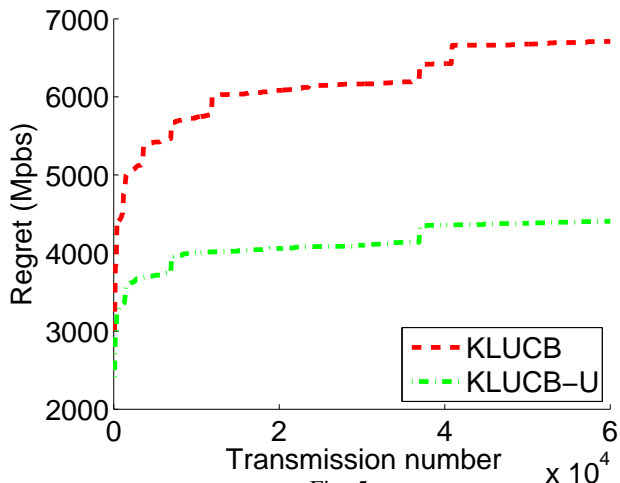


Fig. 5

SIMULATION: REGRET OF DIFFERENT DECISION RULES AS A FUNCTION OF TIME IN A STATIONARY ENVIRONMENT.

in VIII-A, we present the best pair as a function of time and the throughputs of different algorithms in Fig.6. Again KL-UCB-U beats KL-UCB.

For both the test-bed and simulation, the performance of KL-UCB-U is rather impressive: its throughput is at least 95% of that of the Oracle, without knowing the throughputs of the various (channel,rate) pairs beforehand. This shows that given a good decision rule, the selection of channel and rate can be done solely based on ACK/NACK feedback with excellent performance. This is critical for real-world systems because feedback of channel measurements is problematic in practice both in terms of delay and overhead.

So far, in non-stationary environments, the packet successful transmission probabilities were evolving slowly.

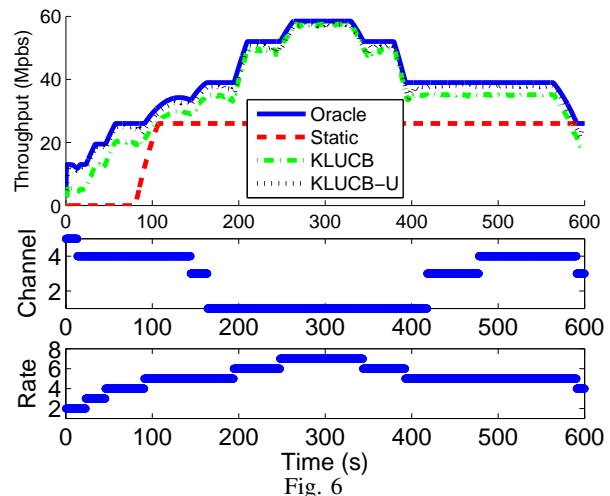


Fig. 6

SIMULATION: THROUGHPUTS OF THE VARIOUS ALGORITHMS (ABOVE) AND THE BEST PAIR  $(c^*, k^*)$  (BELOW) AS A FUNCTION OF TIME IN A NON-STATIONARY ENVIRONMENT – LOW VARIATION SPEED.

Speed	$\times 1$	$\times 20$	$\times 100$
Static	52 %	45 %	43 %
KL-UCB	90 %	83 %	57 %
KL-UCB-U	96 %	91 %	79 %
Oracle	100 %	100 %	100 %

Fig. 7

IMPACT OF THE SPEED OF VARIATION OF THE SUCCESSFUL TRANSMISSION PROBABILITIES ON PERFORMANCE IN A NON-STATIONARY ENVIRONMENT.

Next we vary the speed at which they evolve, by artificially accelerating our traces by a factor 20 and 100. Results are presented in Fig. 7. At all speeds, KL-UCB-U beats KL-UCB, and the performance gap between the two algorithms increases with the speed. When the environment changes faster, the performance of KL-UCB becomes poor, as the algorithm needs to explore all (channel, rate) pairs, and cannot track the best pair. On the contrary, KL-UCB-U exploits the structure and explores less, which makes its performance more robust.

## IX. CONCLUSION

In this paper, we have addressed the problem of joint channel and rate adaptation in cognitive radio systems. We have shown that the problem is equivalent to a structured MAB problem, where the structure stems from inherent properties of the throughput as a function of the selected channel and rate. For several assumptions on this structure, we have derived fundamental

performance limits satisfied by any sequential (channel, rate) selection algorithm. For each structure type, we have also proposed algorithms which are either close or achieve these limits. Finally we have assessed the efficiency of the proposed algorithms through trace-driven experiments and simulations. The two key insights from our results are: (a) The channel and rate adaptation problem has a strong structure. This structure can be exploited to devise algorithms whose performance does not depend on the number of available rates, and is close to that of an Oracle algorithm that perfectly knows the packet successful transmission probabilities at any available (channel, rate) pair. (b) There exist readily implementable algorithms which allow almost perfect channel and rate selection without the need of any measurement and explicit feedback of the quality of the various channels.

## REFERENCES

- [1] FCC, "Second memorandum opinion and order, fcc 10-174," 2010.
- [2] J. Camp and E. Knightly, "Modulation rate adaptation in urban and vehicular environments: cross-layer implementation and experimental evaluation," in *Proceedings of ACM Mobicom*, 2008.
- [3] B. Radunovic, A. Proutiere, D. Gunawardena, and P. Key, "Dynamic channel, rate selection and scheduling for white spaces," in *Proceedings of ACM CoNEXT*, 2011.
- [4] S. Sen, N. Santhapuri, R. R. Choudhury, and S. Nelakuditi, "Accurate: Constellation based rate estimation in wireless networks," in *Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'10, 2010.
- [5] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Predictable 802.11 packet delivery from wireless channel measurements," in *Proceedings of the ACM SIGCOMM 2010 Conference*, ser. SIGCOMM '10, 2010, pp. 159–170.
- [6] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–2, 1985.
- [7] H. Kim and K. G. Shin, "In-band spectrum sensing in cognitive radio networks: Energy detection or feature detection?" in *Proceedings of the 14th ACM International Conference on Mobile Computing and Networking*, ser. MobiCom '08, 2008.
- [8] A. Motamedi and A. Bahai, "Optimal channel selection for spectrum-agile low-power wireless packet switched networks in unlicensed band," *EURASIP J. Wireless Comm. and Networking*, 2008.
- [9] K. Liu, Q. Zhao, and B. Krishnamachari, "Dynamic multichannel access with imperfect channel state detection," *IEEE Trans. Signal Process.*, pp. 2795–2808, 2010.
- [10] L. Lai, H. Jiang, and H. V. Poor, "Medium access in cognitive radio networks: A competitive multi-armed bandit framework," in *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*. IEEE, 2008, pp. 98–102.
- [11] L. Lai, H. El Gamal, H. Jiang, and H. Poor, "Cognitive medium access: Exploration, exploitation, and competition," *Mobile Computing, IEEE Transactions on*, vol. 10, no. 2, pp. 239–253, 2011.
- [12] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive mac for opportunistic spectrum access in ad hoc networks: A pomdp framework," *Selected Areas in Communications, IEEE Journal on*, vol. 25, no. 3, pp. 589–600, 2007.
- [13] S. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *Information Theory, IEEE Transactions on*, vol. 55, no. 9, pp. 4040–4050, 2009.
- [14] J. Bicket, "Bit-rate selection in wireless networks," Ph.D. dissertation, Massachusetts Institute of Technology, 2005.
- [15] S. H. Y. Wong, H. Yang, S. Lu, and V. Bharghavan, "Robust rate adaptation for 802.11 wireless networks," in *Proceedings of the 12th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '06, 2006.
- [16] L. Deek, E. Garcia-Villegas, E. Belding, S.-J. Lee, and K. Almeroth, "Joint rate and channel width adaptation in 802.11 mimo wireless networks," in *Proceedings of IEEE Secon*, 2013.
- [17] C. Tekin and M. Liu, "Online learning in opportunistic spectrum access: A restless bandit approach," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 2462–2470.
- [18] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–222, 2012.
- [19] L. Kocsis and C. Szepesvári, "Discounted ucb," in *Proceedings of the 24th PASCAL Challenges Workshop*, 2006.
- [20] J. Y. Yu and S. Mannor, "Piecewise-stationary bandit problems with side observations," in *ICML*, 2009, p. 148.
- [21] A. Garivier and O. Cappé, "The kl-ucb algorithm for bounded stochastic bandits and beyond," in *Proceedings of Conference On Learning Theory COLT*, 2011.
- [22] J. Y. Yu and S. Mannor, "Unimodal bandits," in *Proceedings of ICML*, 2011.
- [23] R. Combes and A. Proutiere, "Unimodal bandits: Regret lower bounds and optimal algorithms," in *Proceedings of ICML*, 2014.
- [24] A. Garivier and E. Moulines, "On upper-confidence bound policies for non-stationary bandit problems," 2008, arXiv e-print. <http://arxiv.org/abs/0805.3415>.
- [25] S. Sen, B. Radunovic, J. Lee, and K. H. Kim, "Cspy: Finding the best quality channel without probing," in *Proceedings of ACM MOBICOM*, 2013.
- [26] T. L. Graves and T. L. Lai, "Asymptotically efficient adaptive choice of control laws in controlled markov chains," *SIAM J. Control and Optimization*, vol. 35, no. 3, pp. 715–743, 1997.

## APPENDIX

### PROOF OF THEOREM 2

We derive here the regret lower bounds for the MAB problem ( $P_U$ ). To this aim, we apply the techniques used by Graves and Lai [26] to investigate efficient adaptive decision rules in controlled Markov chains. We recall here their general framework. Consider a controlled Markov chain  $(X_t)_{t \geq 0}$  on a finite state space  $\mathcal{S}$  with a control set  $U$ . The transition probabilities given control  $u \in U$  are parametrized by  $\theta$  taking values in a compact metric space  $\Theta$ : the probability to move from state  $x$  to state  $y$  given the control  $u$  and the parameter  $\theta$  is  $p(x, y; u, \theta)$ . The parameter  $\theta$  is not known. The decision maker is provided with a finite set of stationary control laws  $G = \{g_1, \dots, g_K\}$  where each control law  $g_j$  is a mapping from  $\mathcal{S}$  to  $U$ : when control law  $g_j$  is

applied in state  $x$ , the applied control is  $u = g_j(x)$ . It is assumed that if the decision maker always selects the same control law  $g$  the Markov chain is then irreducible with stationary distribution  $\pi_\theta^g$ . Now the reward obtained when applying control  $u$  in state  $x$  is denoted by  $r(x, u)$ , so that the expected reward achieved under control law  $g$  is:  $\mu_\theta(g) = \sum_x r(x, g(x))\pi_\theta^g(x)$ . There is an optimal control law given  $\theta$  whose expected reward is denoted  $\mu_\theta^* \in \arg \max_{g \in G} \mu_\theta(g)$ . Now the objective of the decision maker is to sequentially control laws so as to maximize the expected reward up to a given time horizon  $T$ . As for MAB problems, the performance of a decision scheme can be quantified through the notion of regret which compares the expected reward to that obtained by always applying the optimal control law.

We now apply the above framework to our MAB problem. For  $(P_U)$ , for all  $c$ , the parameter  $\theta_c$  takes values in  $\mathcal{T} \cap \mathcal{U}$ . The Markov chain has values in  $\mathcal{S} = \{0, r_1, \dots, r_K\}$ . The set of control laws is  $G = \{1, \dots, C\} \times \{1, \dots, K\}$ . These laws are constant, in the sense that the control applied by control law  $(c, k)$  does not depend on the state of the Markov chain, and corresponds to selecting (channel, rate) pair  $(c, k)$ . The transition probabilities are given as follows: for all  $x, y \in \mathcal{S}$ ,

$$p(x, y; (c, k), \theta) = p(y; (c, k), \theta) = \begin{cases} \theta_{ck}, & \text{if } y = r_k, \\ 1 - \theta_{ck}, & \text{if } y = 0. \end{cases}$$

Finally, the reward  $r(x, (c, k))$  does not depend on the state and is equal to  $r_k \theta_{ck}$ , which is also the expected reward obtained by always using control law  $(c, k)$ .

We now fix  $\theta$ :  $\forall c, \theta_c \in \mathcal{T} \cap \mathcal{U}$ . Define  $I^{(c,k)}(\theta, \lambda) = I(\theta_{ck}, \lambda_{ck})$  for any  $(c, k)$ . Further define the set  $B(\theta)$  consisting of all *bad* parameters  $\lambda$ :  $\forall c, \lambda_c \in \mathcal{T} \cap \mathcal{U}$  such that  $(c^*, k^*)$  is not optimal under parameter  $\lambda$ , but which are statistically *indistinguishable* from  $\theta$ :

$$B(\theta) = \{\lambda : \forall c, \lambda_c \in \mathcal{T} \cap \mathcal{U}, \\ \lambda_{c^*k^*} = \theta_{c^*k^*}, \max_{(c,k)} r_k \lambda_{ck} > \mu^*\},$$

$B(\theta)$  can be written as the union of sets  $B_{ck}(\theta)$  defined as:

$$B_{ck}(\theta) = \{\lambda \in B(\theta) : r_k \lambda_{ck} > r_{k^*} \lambda_{c^*k^*}\}.$$

Note that  $B_{ck}(\theta) = \emptyset$  if  $r_k < r_{k^*} \theta_{c^*k^*}$ , hence if  $k \notin N$ .

By applying Theorem 1 in [26], we know that  $c_U(\theta)$  is the minimal value of the following LP:

$$\min \sum_{c,k} \alpha_{ck} (\mu^* - r_k \theta_{ck}) \quad (6)$$

$$\text{s.t.} \quad \inf_{\lambda \in B(\theta)} \sum_{(c,k) \neq (c^*, k^*)} \alpha_{ck} I^{(c,k)}(\theta, \lambda) \geq 1, \quad (7)$$

$$\alpha_{ck} \geq 0, \quad \forall (c, k). \quad (8)$$

Next we detail the constraints (7). These constraints are equivalent to:

$$\inf_{\lambda \in B_{c^*k}(\theta)} \sum_{(c,l) \neq (c^*, k^*)} \alpha_{cl} I^{(c,l)}(\theta, \lambda) \geq 1, \quad \forall k \neq k^* \quad (9)$$

$$\inf_{\lambda \in B_{ck^*}(\theta)} \sum_{(c',l) \neq (c^*, k^*)} \alpha_{c'l} I^{(c',l)}(\theta, \lambda) \geq 1, \quad \forall c \neq c^* \quad (10)$$

$$\inf_{\lambda \in B_{ck}(\theta)} \sum_{(c',l) \neq (c^*, k^*)} \alpha_{c'l} I^{(c',l)}(\theta, \lambda) \geq 1, \quad \forall c \neq c^*, \forall k \neq k^*. \quad (11)$$

Constraint (9). We prove that (9) is equivalent to:

$$\min_{k \in M} \alpha_{c^*k} I(\theta_{c^*k}, \frac{\mu^*}{r_k}) \geq 1. \quad (12)$$

Observe that if  $k < k_0$  (i.e., if  $k \notin N$ ), then  $B_{c^*k}(\theta) = \emptyset$ . Let  $k \in N$  with  $k \neq k^*$ . Without loss of generality assume that  $k > k^*$ . We prove that:

$$\inf_{\lambda \in B_{c^*k}(\theta)} \sum_{(c,l) \neq (c^*, k^*)} \alpha_{cl} I^{(cl)}(\theta, \lambda) = \sum_{l=k^*+1}^k \alpha_{c^*l} I(\theta_{c^*l}, \frac{\mu^*}{r_l}). \quad (13)$$

This is due to the fact that we can always choose  $\lambda_{cl} = \theta_{cl}$  for all  $c \neq c^*$ , and to the following two observations:

- for all  $\lambda \in B_{c^*k}(\theta)$ , we have  $\lambda_{c^*k^*} r_{k^*} = \theta_{c^*k^*} r_{k^*}$  and  $\lambda_{c^*k} r_k > \lambda_{c^*k^*} r_{k^*}$ , which using the unimodality of  $\lambda$ , implies that for any  $l \in \{k^*, \dots, k\}$ ,  $\lambda_{c^*l} r_l \geq \theta_{c^*k^*} r_{k^*}$ . Hence:

$$\sum_{l \neq k^*} \alpha_{c^*l} I^{(c^*,l)}(\theta, \lambda) \geq \sum_{l=k^*+1}^k \alpha_{c^*l} I(\theta_{c^*l}, \frac{\mu^*}{r_l}).$$

- For  $\epsilon > 0$ , define  $\lambda_\epsilon$  as follows: for all  $l \in \{k^*, \dots, k\}$ ,  $\lambda_{c^*l} = (1 + (l - k^*)\epsilon) \frac{\mu^*}{r_l}$ , and for all  $l \notin \{k^*, \dots, k\}$ ,  $\lambda_{c^*l} = \theta_{c^*l}$ . By construction,  $\lambda_\epsilon \in B_{c^*k}(\theta)$ , and

$$\lim_{\epsilon \rightarrow 0} \sum_{l \neq k^*} \alpha_{c^*l} I^{(c^*,l)}(\theta, \lambda_\epsilon) = \sum_{l=k^*+1}^k \alpha_{c^*l} I(\theta_{c^*l}, \frac{\mu^*}{r_l}).$$

From (13), we deduce that constraints (7) are equivalent to (12) (indeed, only the constraints related to  $k \in M$  are really active, and for  $k \in M$ , (7) is equivalent to  $\alpha_{c^*k} I(\theta_{c^*k}, \frac{\mu^*}{r_k}) \geq 1$ ).

Constraint (10). Note that if  $k_c^* < k_0$ , then  $B_{ck_c^*}(\theta) = \emptyset$ . Assume that  $k_c^* \geq k_0$ . When  $\lambda \in B_{ck_c^*}(\theta)$ , the optimal (channel, rate) pair under  $\lambda$  is  $(c, k_c^*)$ . This implies that  $r_{k_c^*} \lambda_{ck_c^*} \geq \mu^*$ , and so:

$$\sum_{(c',l) \neq (c^*, k^*)} \alpha_{c'l} I^{(c',l)}(\theta, \lambda) \geq \alpha_{ck_c^*} I(\theta_{ck_c^*}, \frac{\mu^*}{r_{k_c^*}}).$$

Now select  $\lambda_\epsilon$  as follows:  $\lambda_{c'k} = \theta_{c'k}$  for all  $(c', k) \neq$

$(c, k_c^*)$ , and  $\lambda_{ck_c^*} = \mu^*/r_{k_c^*} + \epsilon$ . Then  $\lambda_\epsilon \in B_{ck_c^*}(\theta)$  for all  $\epsilon > 0$ , and

$$\lim_{\epsilon \rightarrow 0} \sum_{(c', l) \neq (c^*, k^*)} \alpha_{c'l} I^{(c', l)}(\theta, \lambda) = \alpha_{ck_c^*} I(\theta_{ck_c^*}, \frac{\mu^*}{r_{k_c^*}}).$$

We conclude that (10) is equivalent to:

$$\forall c \neq c^*, \alpha_{ck_c^*} I(\theta_{ck_c^*}, \frac{\mu^*}{r_{k_c^*}}) \geq 1_{k_c^* \geq k_0}.$$

Constraint (11). For  $k < k_0$ ,  $B_{ck}(\theta) = \emptyset$ . Assume that  $k \geq k_0$  and  $k \neq k_c^*$ . Then  $\sum_{(c', l) \neq (c^*, k^*)} \alpha_{c'l} I^{(c', l)}(\theta, \lambda)$  is minimized over  $B_{ck}(\theta)$  when for all  $c' \neq c$  and all  $k$ ,  $\lambda_{c'k} = \theta_{c'k}$ , and is actually equal to:  $\inf_{\lambda_c \in C_k} \sum_l \alpha_{cl} I(\theta_{cl}, \lambda_{cl})$ . Unfortunately, the above optimization problem cannot be further reduced.  $\square$

### PROOF OF THEOREM 3

For  $\lambda$ :  $\forall c, \theta_c \in \mathcal{T} \cap \mathcal{U}$  and  $\alpha \in \mathbb{R}_+^{C \times K}$ , we define:

$$D(\theta, \lambda, \alpha) = \sum_{c, k} \alpha_{ck} I(\theta_{ck}, \lambda_{ck}).$$

As defined previously, the “bad parameter set” is:

$$B(\theta) = \{\lambda \in \mathcal{T} \cap \mathcal{U} : \lambda_{c^*k^*} = \theta_{c^*k^*}, \max_{(c, k)} r_k \lambda_{ck} > \mu^*\}.$$

Further define the set:

$$\mathcal{C} = \{\alpha \in \mathbb{R}_+^{C \times K} : \inf_{\lambda \in B(\theta)} D(\theta, \lambda, \alpha) \geq 1\}.$$

$c_U(\theta)$  in Theorem 2 is the solution to a minimization problem over  $\mathcal{C}$ . An upper bound of  $c_U(\theta)$  is obtained by choosing  $\alpha \in \mathcal{C}$  and by computing the value of the objective function at  $\alpha$  which is  $\sum_{c, k} \alpha_{ck} (\mu^* - \mu_{ck})$ . We prove that if we define  $\alpha$  as:

- $\alpha_{c^*k} = 1/I(\theta_{c^*k}, \frac{\mu^*}{r_k})$  if  $k \in M$ ,
- $\alpha_{ck_c^*} = (\min\{I(\theta_{ck_c^*}, \frac{\mu^*}{r_{k_c^*}}), I(\theta_{ck_c^*}, \theta_{ck_c^*} - \frac{\delta_c}{r_{k_c^*}})\})^{-1}$  if  $c \neq c^*$ ,
- $\alpha_{ck} = 1/I(\theta_{ck}, \theta_{ck} + \frac{\delta_c}{r_k})$  if  $c \neq c^*$  and  $k \in M_c$ ,
- $\alpha_{ck} = 0$  if  $c \neq c^*$  and  $k \notin M_c \cup \{k_c^*\}$ .

then  $\alpha \in \mathcal{C}$ . To do so, we use the following decomposition:  $B(\theta) = \cup_{(c, k) \neq (c^*, k^*)} B_{ck}(\theta)$  where

$$B_{ck}(\theta) = \{\lambda \in B(\theta) : (c, k) \in \arg \max_{(c', k')} r_{k'} \lambda_{c'k'}\}.$$

(i) If  $\lambda \in \cup_{k \neq k^*} B_{c^*k}(\theta)$ . Since  $\lambda \in \cup_{k \neq k^*} B_{c^*k}(\theta)$ , we

have  $\theta_{c^*k^*} = \lambda_{c^*k^*}$ . Since  $k \mapsto r_k \lambda_{c^*k}$  is unimodal, and  $k^* \notin \arg \max_k r_k \lambda_{c^*k}$ , then there must exist a neighbour  $k'$  of  $k^*$  such that  $r_{k'} \lambda_{c^*k'} \geq r_{k^*} \lambda_{c^*k^*} = r_{k^*} \theta_{c^*k^*} = \mu^*$ . Hence  $\lambda_{c^*k'} \geq \mu^*/r_{k'}$ . Using the monotonicity of the

KL divergence:

$$\begin{aligned} D(\theta, \lambda, \alpha) &\geq \alpha_{c^*k'} I(\theta_{c^*k'}, \lambda_{c^*k'}) \\ &\geq \alpha_{c^*k'} I(\theta_{c^*k'}, \mu^*/r_{k'}) \\ &\geq 1. \end{aligned}$$

(ii) If  $\lambda \in \cup_k B_{ck}(\theta)$ ,  $c \neq c^*$ . Under parameter  $\lambda$ , let

$\tilde{k} = \arg \max_k r_k \lambda_{ck}$  be the optimal rate for channel  $c$ . We further consider two cases depending on whether  $\tilde{k}$  is equal to  $k_c^*$ .

Case (a):  $\tilde{k} = k_c^*$ . Then  $\lambda \in B_{ck_c^*}(\theta)$ , and we have  $r_{k_c^*} \lambda_{ck_c^*} \geq \mu^*$ . Hence:

$$\begin{aligned} D(\theta, \lambda, \alpha) &\geq \alpha_{ck_c^*} I(\theta_{ck_c^*}, \lambda_{ck_c^*}) \\ &\geq \alpha_{ck_c^*} I(\theta_{ck_c^*}, \mu^*/r_{k_c^*}) \\ &\geq 1. \end{aligned}$$

Case (b):  $\tilde{k} \neq k_c^*$ . Since  $k \mapsto r_k \lambda_{ck}$  is unimodal and  $k_c^* \neq \arg \max_k r_k \lambda_{ck}$ , there must exist a neighbour  $k'$  of  $k_c^*$  such that  $r_{k'} \lambda_{ck'} \geq r_{k_c^*} \lambda_{ck_c^*}$ . Since  $k \mapsto r_k \theta_{ck}$  is unimodal and  $k_c^* = \arg \max_k r_k \theta_{ck}$ , we have  $r_{k_c^*} \theta_{ck_c^*} \geq r_{k'} \theta_{ck'}$ . Therefore:

$$\begin{aligned} \max(r_{k_c^*} |\lambda_{ck_c^*} - \theta_{ck_c^*}|, r_{k'} |\lambda_{ck'} - \theta_{ck'}|) \\ \geq (r_{k_c^*} \theta_{ck_c^*} - r_{k'} \theta_{ck'})/2 \geq \delta_c. \end{aligned}$$

To establish the above inequality, we have used the fact that for all  $a, b > 0$  and for all  $x \in \mathbb{R}$ ,  $\max(|x|, |x + a + b|) \geq (a + b)/2$ , and have applied this result for  $x = r_{k_c^*} (\lambda_{ck_c^*} - \theta_{ck_c^*})$ ,  $a = r_{k'} \lambda_{ck'} - r_{k_c^*} \lambda_{ck_c^*}$ , and  $b = r_{k_c^*} \theta_{ck_c^*} - r_{k'} \theta_{ck'}$ . We have shown that:

- either (b1)  $r_{k_c^*} |\lambda_{ck_c^*} - \theta_{ck_c^*}| \geq \delta_c$ ;  
or (b2)  $r_{k'} |\lambda_{ck'} - \theta_{ck'}| \geq \delta_c$ .

If (b1) holds, then either  $\lambda_{ck_c^*} \leq \theta_{ck_c^*} - \delta_c/r_{k_c^*}$  or  $\lambda_{ck_c^*} \geq \theta_{ck_c^*} + \delta_c/r_{k_c^*}$ . In the latter case, we have:

$$\begin{aligned} \lambda_{ck'} &\geq \frac{r_{k_c^*}}{r_{k'}} \lambda_{ck_c^*} \geq \frac{r_{k_c^*}}{r_{k'}} \theta_{ck_c^*} + \frac{\delta_c}{r_{k'}} \\ &\geq \theta_{ck'} + \frac{\delta_c}{r_{k'}}. \end{aligned}$$

If (b2) holds, then either  $\lambda_{ck'} \geq \theta_{ck'} + \delta_c/r_{k'}$ , or  $\lambda_{ck'} \leq \theta_{ck'} - \delta_c/r_{k'}$ . In the latter case, we have:

$$\begin{aligned} \lambda_{ck_c^*} &\leq \frac{r_{k'}}{r_{k_c^*}} \lambda_{ck'} \leq \frac{r_{k'}}{r_{k_c^*}} \theta_{ck'} - \frac{\delta_c}{r_{k_c^*}} \\ &\leq \theta_{ck_c^*} - \frac{\delta_c}{r_{k_c^*}}. \end{aligned}$$

In both cases (b1) and (b2), we have proved that either

$\lambda_{ck_c^*} \leq \theta_{ck_c^*} - \delta_c/r_{k_c^*}$  or  $\lambda_{ck'} \geq \theta_{ck'} + \delta_c/r_{k'}$ . Finally:

$$\begin{aligned} D(\theta, \lambda, \alpha) &\geq \alpha_{ck_c^*} I(\theta_{ck_c^*}, \lambda_{ck_c^*}) + \alpha_{ck'} I(\theta_{ck'}, \lambda_{ck'}) \\ &\geq \max\{\alpha_{ck_c^*} I(\theta_{ck_c^*}, \theta_{ck_c^*} - \frac{\delta_c}{r_{k_c^*}}), \\ &\quad \alpha_{ck'} I(\theta_{ck'}, \theta_{ck'} + \frac{\delta_c}{r_{k'}})\} \\ &\geq 1. \end{aligned}$$

We have proved that  $\inf_{\lambda \in B(\theta)} D(\theta, \lambda, \alpha) \geq 1$ , and thus  $\alpha \in \mathcal{C}$ . Now for our choice of  $\alpha$ , the value of the objective function of the optimization problem in Theorem 2 is  $c'_u(\theta)$ . We conclude that  $c_U(\theta) \leq c'_U(\theta)$ .  $\square$

### PROOF OF THEOREM 6

Let  $\pi = \text{CRS-T}$ . We denote by  $r^\pi(T)$  the sample path regret under CRS-T up to time  $T$ :

$$r^\pi(T) = \mu^* - \sum_{c,k} \mu_{ck} t_{ck}(T),$$

so that  $R^\pi(T) = \mathbb{E}[r^\pi(T)]$ .

Define  $\Delta = \min_{(c,k),(c',k'): \mu_{ck} \neq \mu_{c'k'}} |\mu_{ck} - \mu_{c'k'}|$  the minimal separation between two decisions. The proof consists in two steps. In the first step, we provide an upper bound of the sample path regret. In the second step, we establish that  $\limsup_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} < \infty$ , and apply the dominated convergence theorem to conclude.

Step 1. Fix a sample path (for the successive rewards obtained under CRS-T). By design of CRS-T,  $t_{ck}(n) \rightarrow \infty$ ,  $n \rightarrow \infty$  a.s. (since when  $t_{ck}(n) < \log(\log(n))$ , decision  $(c, k)$  is taken). Hence  $\hat{\mu}_{ck}(n) \rightarrow \mu_{ck}$  a.s. by the law of large numbers. Let  $\delta < \Delta/2$  – the choice of  $\delta$  will be made more precise later. There exists  $n_0 \geq 1$  (that depends on  $\delta$ ) such that for all  $n \geq n_0$ ,  $|\hat{\mu}_{ck}(n) - \mu_{ck}| \leq \delta$  for all  $(c, k)$ . As a consequence, for  $n \geq n_0$ , for any channel  $c$ , the leader  $l_c(n)$  corresponds to the best rate  $k_c^*$ . Also observe that for  $n \geq n_0$ , for all  $k \in N(k_c^*)$ ,  $\delta_{ck}(n) > 0$ . Hence we cannot have  $U_c(n) = -1$ , and  $U_c(n) \in \{0, 1\}$ .

Next, we provide an upper bound of  $t_{ck}(n)$  for all  $(c, k) \neq (c^*, k^*)$ .

(a) For any channel  $c$ , let  $k \notin N(k_c^*)$ . When  $n \geq n_0$ , for all  $n' \in \{n_0, \dots, n\}$ , decision  $(c, k)$  is selected for the  $(n' + 1)$ -th transmission only if  $t_{ck}(n') < \log(\log(n'))$ . We simply deduce that:

$$t_{ck}(n) \leq n_0 + \log(\log(n)).$$

(b) For any channel  $c$ , let  $k \in N(k_c^*)$ . Let  $n \geq n_0$ , and  $n' \in \{n_0, \dots, n\}$ . To ease the presentation, we define

the following quantity:

$$\tau_{ck}(\delta) = \frac{\alpha(r_k^2 + r_{k_c^*}^2)}{2(\mu_{ck_c^*} - \mu_{ck} - 2\delta)^2}.$$

Assume that:

$$t_{ck}(n') \geq \log(n) \tau_{ck}(\delta).$$

There are two possibilities: if  $t_{ck_c^*}(n') < \log(n) \tau_{ck}(\delta)$ , then  $(c, k)$  is not selected for transmission at  $n''$  except if  $t_{ck}(n'') < \log(\log(n''))$ , since the selected decision  $(c, k)(n'')$  must be included in  $\arg \min_{k' \in \mathcal{V}_c(n'')}(n'') t_{c(n'')k'}(n'')$  (remember that if  $U_c(n'') = 1$  then the index of  $(c, k)$  is equal to 0, and  $(c, k)$  is not selected). If  $t_{ck_c^*}(n') \geq \log(n) \tau_{ck}(\delta)$ , then:

$$\min(t_{ck}(n'), t_{ck_c^*}(n')) \geq \log(n) \tau_{ck}(\delta).$$

From there, we deduce that:

$$\begin{aligned} \bar{\delta}_{ck}(n'') &\leq \sqrt{\alpha \log(n'') \frac{r_k^2 + r_{k_c^*}^2}{2 \log(n) \tau_{ck}(\delta)}} \\ &\leq \mu_{ck_c^*} - \mu_{ck} - 2\delta. \end{aligned}$$

Hence  $\delta_{ck}(n'') - \bar{\delta}_{ck}(n'') > 0$  for all  $n'' \in \{n', \dots, n\}$ . Thus  $k \notin \mathcal{V}_c(n'')$ , and  $(c, k)$  is not selected at for the transmissions in  $\{n', \dots, n\}$ , except if  $t_{ck}(n'') < \log(\log(n''))$ . We deduce that:

$$t_{ck}(n) \leq n_0 + \log(\log(n)) + \log(n) \tau_{ck}(\delta).$$

(c) If  $c \neq c^*$ ,  $k = k_c^*$ , using the same reasoning as above:

$$\begin{aligned} t_{ck}(n) &\leq n_0 + \log(\log(n)) \\ &\quad + \log(n) \max \left( \frac{1}{I(\frac{\mu_{ck} + \delta}{r_k}, \frac{\mu^* - \delta}{r_k})}, \max_{k \in N(k_c^*)} \tau_{ck}(\delta) \right). \end{aligned}$$

Now using the continuity of the KL divergence, for any  $\epsilon > 0$ , we can choose  $\delta > 0$  such that for all  $n \geq n_0$  (note that  $n_0$  depends on  $\epsilon$ ), almost surely:

For all  $c$ , and all  $k \notin N(k_c^*)$ ,

$$t_{ck}(n) \leq n_0 + \log(\log(n));$$

for all  $c$ , and all  $k \in N(k_c^*)$ ,

$$t_{ck}(n) \leq n_0 + \log(\log(n)) + \log(n)(\tau_{ck} + \epsilon);$$

for all  $c \neq c^*$ ,  $k = k_c^*$ ,

$$\begin{aligned} t_{ck}(n) &\leq n_0 + \log(\log(n)) \\ &\quad + \log(n) \left[ \max \left( \frac{1}{I(\frac{\mu_{ck}}{r_k}, \frac{\mu^*}{r_k})}, \max_{k \in N(k_c^*)} \tau_{ck} \right) + \epsilon \right]. \end{aligned}$$



We conclude that, almost surely:

$$\limsup_{T \rightarrow \infty} \frac{r^\pi(T)}{\log(T)} \leq c^{\text{CRS-T}}(\theta).$$

Step 2. The following lemma ensures that the average regret is bounded:

*Lemma 1:* Under algorithm CRS-T, the regret is upper-bounded uniformly by:

$$\sup_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} < +\infty$$

Combining the results of step 1 and of the above lemma, we complete the proof of Theorem 6 by simply applying the dominated convergence theorem.  $\square$

*Proof of Lemma 1.* We decompose the set of instants  $\{1, \dots, T\}$  at which we may play a sub-optimal (channel, rate) pair as follows. We define the following sets:

- $A = \{1 \leq n \leq T : \min_{ck} t_{ck}(n) < \log(\log(n))\}$ , the set of instants at which there exists a decision which has not been selected  $\log(\log(n))$ .
- $B = \cup_{ck} B_{ck}$ , where  $B_{ck} = \{1 \leq n \leq T : c(n) = c, l_c(n) = k, U_c(n) = 0\}$  is the set of instants at which  $k$  is the leader for channel  $c$ , channel  $c$  is selected and there are not enough samples to determine whether  $k$  is better than all its neighbours.
- $D = \cup_c D_c^1 \cup D_c^{-1}$ , where  $D_c^1 = \{1 \leq n \leq T : l_c(n) \neq k_c^*, U_c(n) = 1\}$ ,  $D_c^{-1} = \{1 \leq n \leq T : l_c(n) = k_c^*, U_c(n) = -1\}$ .  $D_c^1 \cup D_c^{-1}$  are the set of the instants at which the test fails for channel  $c$ : Either we believe that  $l_c(n)$  is  $k_c^*$  when it is not the case ( $D_c^1$ ), or we believe that  $l_c(n)$  is different from  $k_c^*$  when the two are equal ( $D_c^{-1}$ ).
- $E = \cup_{(c,k) \neq (c^*, k^*)} E_{ck}$ , where  $E_{ck} = \{1 \leq n \leq T, q_{ck}(n) < \mu_{ck}\}$  is the set of instants at which for the pair  $(c, k)$ , the upper confidence bound  $\bar{b}_{ck}(n)$  underestimates the actual average reward  $\mu_{ck}$ .
- $F = \cup_{(c,k) \neq (c^*, k^*)} F_{ck}$ , where  $F_{ck} = \{1 \leq n \leq T, \mathcal{U}(n) = \emptyset, n \notin A \cup D \cup E, (c, k)(n) = (c, k)\}$ .  $F$  is the set of instants at which the test has returned a correct answer for all channels, the average rewards are not underestimated, and yet the sub-optimal rate  $(c, k)$  is selected.

We have:

$$R^\pi(T) \leq r_K(\mathbb{E}[|A|] + \mathbb{E}[|B|] + \mathbb{E}[|D|] + \mathbb{E}[|E|] + \mathbb{E}[|F|]).$$

Next we bound the expected size of sets  $A, B, D, E$

and  $F$ .

Upper bound of  $\mathbb{E}[|A|]$ . By design of CRS-T, if  $n \in A$  then the pair  $(c, k)$  which has been the least tried is selected, so that  $\mathbb{E}[A] \leq CK \log(\log(T))$ .

Upper bound of  $\mathbb{E}[|B|]$ . Let  $n \in B_{ck}$ , and define  $s = \sum_{n'=1}^n \mathbf{1}\{n' \in B_{ck}\}$ . Then for all  $k' \in N(k) \cup \{k\}$ ,  $t_{ck'}(n) \geq s/(\gamma + 1)$  since when  $n \in B_{ck}$ , the selected rate is the rate which has been played the least among  $k$  and its neighbours on channel  $c$ . We recall that  $\inf_{k' \in N(k)} |\mu_{ck} - \mu_{ck'}| > \Delta$ . Since  $n \in B_{ck}$ , we have  $U_c(n) = 0$  and  $l_c(n) = k$  so there must exist  $k' \in N(k)$  such that  $\delta_{ck'}(n) \in [-\bar{\delta}_{ck'}(n), \bar{\delta}_{ck'}(n)]$ . Define

$$s_0 = 4\alpha \log(T) r_K^2(\gamma + 1) \Delta^{-2}.$$

Let  $s > s_0$ . Then we have that  $\bar{\delta}_{ck'}(n) \leq \Delta/2$ . First consider the case where  $\mu_{ck} > \mu_{ck'}$ . We have  $\mu_{ck} - \mu_{ck'} \geq \Delta \geq 2\bar{\delta}_{ck'}(n)$ , and  $\delta_{ck'}(n) - \bar{\delta}_{ck'}(n) \leq 0$ . Therefore we have:

$$\delta_{ck'}(n) \leq (\mu_{ck} - \mu_{ck'}) - \bar{\delta}_{ck'}(n).$$

Applying the second inequality of Lemma 3 (presented at the end of the proof), we deduce:

$$\mathbb{P}[\delta_{ck'}(n) \leq (\mu_{ck} - \mu_{ck'}) - \bar{\delta}_{ck'}(n)] \leq n^{-\alpha}.$$

The case  $\mu_{ck} < \mu_{ck'}$  is treated similarly, and we have proved that:

$$\mathbb{P}[n \in B_{ck}, s \geq s_0] \leq C_0 n^{-\alpha},$$

for some constant  $C_0$ . Using the fact that  $\sum_{n=1}^T \mathbf{1}\{n \in B_{ck}, s \leq s_0\} \leq s_0$ :

$$\begin{aligned} \mathbb{E}[|B_{ck}|] &\leq s_0 + \sum_{n=1}^T \mathbb{P}[n \in B_{ck}, s \geq s_0] \\ &\leq s_0 + \sum_{n=1}^T C_0 n^{-\alpha} = O(\log(T)). \end{aligned}$$

Hence  $\mathbb{E}[|B|] \leq O(\log(T))$ .

Upper bound of  $\mathbb{E}[|D|]$ . Let  $n \in D_c^1$ . Since  $l_c(n) \neq k_c^*$ , and  $U_c(n) = 1$ , there must exist a couple  $(k, k')$  such that  $k = l_c(n)$ ,  $k' \in N(k)$ ,  $\mu_{kc} - \mu_{k'c} < 0$  and  $\delta_{ck'}(n) - \bar{\delta}_{ck'}(n) > 0$ . Applying the first inequality of Lemma 3:

$$\mathbb{P}[\delta_{ck'}(n) - \bar{\delta}_{ck'}(n) > 0] \leq$$

$$\mathbb{P}[\delta_{ck'}(n) \geq (\mu_{kc} - \mu_{k'c}) + \bar{\delta}_{ck'}(n)] \leq n^{-\alpha}.$$

Applying the union bound, we get  $\mathbb{P}[n \in D_c^1] \leq C_1 n^{-\alpha}$  for some constant  $C_1$ . By symmetry we have that  $\mathbb{P}[n \in D_c^{-1}] \leq C_1 n^{-\alpha}$ . Therefore  $\mathbb{E}[|D_c|] = O(1)$  and  $\mathbb{E}[|D|] = O(1)$ .

Upper bound of  $\mathbb{E}[|E|]$ . In view of Lemma 2 (presented

below),  $\mathbb{E}[|E_{ck}|] = O(\log(\log(T)))$ , so that  $\mathbb{E}[|E|] = O(\log(\log(T)))$ .

Upper bound of  $\mathbb{E}[|F|]$ . Let  $n \in F_{ck}$ . Since  $\mathcal{U}(n) = \emptyset$  and  $n \notin A$ , and  $(c, k)(n) = (c, k)$ , we must have that  $(c, k) \in \arg \max_{ck} b_{ck}(n)$ .

- If  $U_{c^*}(n) = 1$ , then  $l_{c^*}(n) = k^*$  because  $n \notin D$  and so  $b_{c^*k^*}(n) = q_{c^*k^*}(n)$ .
- If  $U_{c^*}(n) = -1$ , then  $b_{c^*k^*}(n) = q_{c^*k^*}(n)$  by design.

Because  $n \notin E$ , we have  $q_{c^*k^*} \geq \mu^*$ . Hence  $b_{ck}(n) \geq \mu^*$ . Following the same reasoning as [21], we prove that:

$$\begin{aligned} \mathbb{E}[|F_{ck}|] &\leq \sum_{n=1}^T \mathbf{1}\{(c, k)(n) = (c, k), b_{ck}(n) \geq \mu^*\} \\ &\leq \frac{\log(T)}{I(\theta_{ck}, \mu^*/r_k)} + o(\log(T)) = O(\log(T)). \end{aligned}$$

So  $\mathbb{E}[|F|] = O(\log(T))$ . This concludes the proof of Lemma 1.  $\square$

The following lemma is proved in [21].

*Lemma 2:* There exists a constant  $\kappa$  such that for all  $(c, k)$ ,  $\mathbb{E}[\sum_{n=1}^T \mathbf{1}\{b_{ck}(n) < \mu_{ck}\}] \leq \kappa \log(\log(T))$ .

Lemma 3 follows from the Azuma-Hoeffding inequality.

*Lemma 3:* We have, for all  $n$  and  $c, c', k, k'$ ,

$$\begin{aligned} \mathbb{P}[\hat{\mu}_{ck}(n) - \hat{\mu}_{c'k'}(n) \geq \mu_{ck} - \mu_{c'k'} \\ + \sqrt{\frac{\alpha \log(n)}{2} \left( \frac{r_k^2}{t_{ck}(n)} + \frac{r_{k'}^2}{t_{c'k'}(n)} \right)}] \leq n^{-\alpha} \end{aligned}$$

and by symmetry:

$$\begin{aligned} \mathbb{P}[\hat{\mu}_{ck}(n) - \hat{\mu}_{c'k'}(n) \leq \mu_{ck} - \mu_{c'k'} \\ - \sqrt{\frac{\alpha \log(n)}{2} \left( \frac{r_k^2}{t_{ck}(n)} + \frac{r_{k'}^2}{t_{c'k'}(n)} \right)}] \leq n^{-\alpha} \end{aligned}$$